Models economètrics avançats amb R

Daniel Liviano Solís
Maria Pujol Jover

PID_00211045

Universitat Oberta de Catalunya

Índex

Introducció 5

Objectius 7

1. Models de regressió dinàmics i multiequacionals 9
1.1. Models de regressió dinàmics 9
1.1.1. Tipus de models dinàmics 9
1.1.2. Anàlisi i interpretació dels models dinàmics 12
1.1.3. Mètodes d'estimació 18
1.2. Models de regressió multiequacionals 24
1.2.1. Hipòtesis bàsiques i formulació general d'un model multiequacional 24
1.2.2. Tipologia de models multiequacionals 25
1.2.3. El problema de la identificació 25
1.2.4. L'estimació dels models multiequacionals 26
1.2.5. Interpretació dels paràmetres del model 28

2. Model lineal generalitzat 29
2.1. Motivació 29
2.2. Models lògit, pròbit i Poisson 30
2.3. Aplicació empírica 32

3. Models amb dades de panel 44
3.1. Introducció 44
3.2. Estimació d'un model de dades de panel 46
3.2.1. Mínims quadrats ordinaris (MCO) 46
3.2.2. Efectes fixos (WG/LSDV) 49
3.2.3. Primeres diferències (FD) 52
3.2.4. Entre grups (BG) 53
3.2.5. Efectes aleatoris (RE/GLS) 53
3.2.6. Coeficients variables 57
3.2.7. Mètode generalitzat dels moments (GMM) 59
3.3. Inferència 60
3.4. Aplicació pràctica amb R 61

Bibliografia 74

Introducció

Aquest mòdul es dedica a l'estudi de models economètrics avançats, l'especificació i estimació dels quals adquireixen certa complexitat. Encara que l'enfocament d'aquest mòdul és eminentment pràctic, i està enfocat al desenvolupament d'aplicacions pràctiques amb R i amb R-Commander, al començament de cada capítol s'ofereix una breu explicació teòrica dels models analitzats. No obstant això, per a entendre els models que exposem aquí, és imprescindible que l'estudiant primer hagi treballat els mòduls teòrics d'econometria, ja que aquest manual no substitueix en cap cas els mòduls teòrics, simplement els complementa.

El primer capítol està dedicat a l'estimació dels models de regressió dinàmics i multiequacionals amb R-Commander. En economia, sovint les relacions entre les variables no es produeixen únicament en el període analitzat, sinó que afecten més d'un període i fins i tot moltes perduren en el temps. Així doncs, quan es treballa amb dades temporals, podem trobar-nos amb freqüència que les relacions deixen de ser estàtiques, passen a ser dinàmiques i donen lloc als models de regressió dinàmics. Com mostrarem a continuació, aquests models tenen les seves pròpies normes d'especificació i d'interpretació associada als paràmetres estimats. També veurem els diferents problemes amb els quals ens podrem trobar a l'hora d'estimar-les i com utilitzar la solució idònia per a solucionar cada un. Finalment, també veurem que molts esdeveniments econòmics s'expliquen per variables que són exògenes i endògenes al mateix temps, ja que s'especifiquen per més d'una equació; són els models multiequacionals. En aquest apartat mostrarem la notació que s'empra per a especificar-los, estimar-los i posteriorment interpretar de manera correcta els paràmetres.

El segon capítol es dedica al model lineal generalitzat (GLM, en les sigles en anglès). Aquest model és una generalització flexible del model de regressió lineal ordinari i permet incloure variables dependents que generen distribucions de l'error diferents d'una distribució normal. Això succeeix, especialment, quan la variable dependent és de naturalesa qualitativa (és a dir, no representa magnituds, sinó diferents atributs o categories), o quan, tot i ser una variable quantitativa, la seva distribució dista de seguir una llei normal. Encara que el GLM inclou una gran multitud de distribucions, en aquesta secció veurem els tres models de regressió més comuns. Això és, per a les variables dependents qualitatives dicotòmiques analitzem els models de regressió lògit i pròbit, i per a les variables de recompte estudiem el model de regressió de Poisson. Encara que el GLM ofereix moltes més possibilitats, aquestes queden fora de l'abast d'aquest manual.

El tercer capítol s'encarrega d'estudiar els models amb dades de panel. El terme *panel de dades* fa referència a un conjunt de dades amb observacions temporals per als mateixos individus, cosa que permet el seguiment d'un mateix individu durant un període de temps. A causa de l'àmplia disponibilitat d'aquest tipus de dades, s'utilitzen en di-

ferents camps, com l'economia, la demografia i les finances, per esmentar-ne alguns. Com que es tracta d'un conjunt de tècniques més aviat complexes, l'anàlisi aplicada es fa exclusivament amb codi R, i va precedit d'una àmplia explicació de la naturalesa i les característiques analítiques d'aquests models, a més de les diferents possibilitats d'especificació i estimació que hi ha.

Objectius

1. Especificar correctament els diferents tipus de models de regressió dinàmics.

2. Saber calcular el multiplicador d'impacte, el multiplicador retardat j períodes i el multiplicador total amb R-Commander.

3. Saber calcular la mitjana i la mediana del retard amb R-Commander i el nombre de períodes que han de transcórrer per assolir un percentatge determinat del canvi que es produirà en la variable endògena davant una modificació en el valor d'una variable explicativa.

4. Saber calcular, utilitzant el mètode de variables instrumentals, els estimadors per mínims quadrats en dues etapes amb R-Commander.

5. Saber escriure la forma estructural, reduïda i final d'un model multiequacional, i interpretar el significat dels seus paràmetres.

6. Identificar les equacions d'un model multiequacional.

7. Ser capaç d'identificar les variables dependents que requereixen de l'estimació d'un model lineal generalitzat.

8. Saber triar, en cada cas, el model més adequat a les característiques de les dades objecte d'estudi.

9. Saber interpretar correctament els coeficients estimats en un model lineal generalitzat.

10. Poder efectuar una anàlisi amb un panel de dades.

11. Saber, en cada cas, quin tipus d'efectes cal incloure en el model: individuals, temporals o tots dos.

12. Triar correctament l'especificació per a cada conjunt de dades de panel.

13. Encertar amb el mètode d'estimació adequat amb un panel de dades, segons si hi ha efectes fixos o aleatoris.

1. Models de regressió dinàmics i multiequacionals

1.1. Models de regressió dinàmics

En els models de regressió tractem d'explicar el comportament d'una variable Y en funció de k variables X. Per tant, sempre que s'alteri el valor d'alguna X es modificarà el valor de Y.

Quan treballem amb sèries temporals no s'ha d'oblidar que la temporalitat estarà sempre implícitament o explícitament en l'especificació del model simplement perquè és un aspecte clau que afecta el valor que pugui prendre la variable que hem d'explicar. Si no plasmem en el model l'efecte temporal, el model estimat no serà l'adequat: els residus podrien estar autocorrelacionats indicant l'omissió de variables rellevants o d'una mala especificació, la bondat de l'ajust no serà tan alta com es podria esperar, etc. La qüestió és que en el model hi ha una relació dinàmica entre les variables que no hem recollit.

En els models dinàmics ens serà molt útil l'operador de retards; recordem com funciona: $Y_{t-1} = LY_t$; i, en general, $Y_{t-j} = L^j Y_t$.

És habitual trobar-nos amb models amb autocorrelació residual d'ordre 1 que capta la relació no contemporània que es dóna en els models amb sèries temporals i que no hem inclòs en l'especificació. Això ens explicita les relacions que hi ha en el terme de pertorbació (entre u_t i u_{t-1}), cosa que ens obligaria a reespecificar el model incloent-hi variables retardades. Les variables retardades no sempre han de formar part del conjunt de variables explicatives de l'especificació inicial, a vegades hem d'incloure com a explicativa la variable endògena retardada. Per tant, els tipus de relacions dinàmiques no sempre són els mateixos, poden sorgir al cap de cert temps, poden estar presents únicament en el període posterior o poden tenir un efecte indefinit però que es va diluint en el temps.

1.1.1. Tipus de models dinàmics

Els principals models dinàmics que podem especificar són:

No oblideu que un model $RD(2)$ sempre serà estable i que els seus paràmetres β s'interpreten com els multiplicadors (retardats o contemporanis) de X sobre Y.

1) Models de retards distribuïts (RD). En un model de retards distribuïts d'ordre s ($RD(s)$) s'explicita la dinamicitat amb la introducció de variables exògenes retardades com a regressors. L'ordre del model de retards distribuïts serà el nombre de retards de les variables exògenes, que pot ser finit o infinit. Per exemple, si una variable Y s'explica per una altra variable X, el model $RD(s)$ podria ser:

$$Y_t = \mu + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_s X_{t-s} + u_t \qquad t = 1, \ldots, T$$

que en funció de l'operador de retards dóna lloc a:

$$Y_t = \mu + \beta_0 X_t + \beta_1 L X_t + \beta_2 L^2 X_t + \ldots + \beta_s L^s X_t + u_t$$

$$= \mu + (\beta_0 + \beta_1 L + \beta_2 L^2 + \ldots + \beta_s L^s) X_t + u_t = \mu + B(L) X_t + u_t$$

Si en lloc de tenir inicialment un MRLS tinguéssim un MRLM, l'especificació d'un model $RD(s_1, s_2, s_3, \ldots, s_k)$ quedaria de la manera següent:

$$Y_t = \mu + B_1(L) X_{1t} + B_2(L) X_{2t} + \ldots + B_k(L) X_{kt} + u_t$$

on $B_j(L)$ és el polinomi s_j associat a la variable X_{ij}.

2) Models autoregressius (AR). Un model autoregressiu d'ordre r és aquell que inclou una sèrie de variables explicatives entre les quals figuren també la mateixa variable endògena retardada com a senyal de la relació no contemporània existent. Aquest model es defineix de la manera següent:

En un model AR d'ordre finit una variació en X tindrà un efecte temporal indefinit sobre Y i el seu impacte total pot ser finit o no; mentre que en un model RD l'impacte és finit i es distribueix en un període temporal determinat.

$$Y_t = \mu + \beta_0 X_t + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_r Y_{t-r} + u_t \qquad t = 1, \ldots, T$$

que en funció de l'operador de retards es transforma en:

$$Y_t - \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2} - \ldots - \alpha_r Y_{t-r} = \mu + \beta_0 X_t + u_t$$

$$(1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_r L^r) Y_t = \mu + \beta_0 X_t + u_t$$

$$A(L) Y_t = \mu + \beta_0 X_t + u_t$$

Especificar un model dinàmic $RD(\infty)$ o $AR(\infty)$ pot provocar problemes perquè ens obliga a estimar infinits paràmetres. Per aquest motiu, una solució consisteix a reespecificar el model en un AR o RD d'ordre finit, respectivament.

3) La hipòtesi de Koyck. Transforma un model de retards distribuïts d'ordre infinit en un model autoregressiu de primer ordre. Partim d'un model $RD(\infty)$:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + u_t$$

i apliquem la hipòtesi de Koyck, que suposa que el valor dels paràmetres **B** disminueix de manera geomètrica ($\beta_j = \beta_0 \delta^j$, $\forall\ j = 0, 1, \ldots$), on $0 < \delta < 1$. De manera que cada vegada que ens allunyem en el temps la influència de X sobre Y va disminuint.

Si tenim en compte la hipòtesi anterior i la substituïm en el model $RD(\infty)$ especificat, arribem a un model $AR(1)$,

$$Y_t = \alpha(1-\delta) + \beta_0 X_t + \delta Y_{t-1} + v_t$$

essent

$$v_t = u_t - \delta u_{t-1}$$

Aquest nou model té l'avantatge que redueix el nombre de paràmetres que s'han d'estimar i els possibles problemes de multicol·linealitat. Per contra, tenim com a explicativa la variable endògena retardada i el nou terme de pertorbació no és esfèric, fet que ens condiciona el mètode d'estimació que s'ha d'utilitzar.

4) Models autoregressius i de retards distribuïts (AD). Un model $AD(r, s)$ es defineix de la manera següent:

$$A(L)Y_t = \mu + B(L)X_t + u_t$$

$$(1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_r L^r)Y_t = \mu + (\beta_0 + \beta_1 L + \beta_2 L^2 + \ldots + \beta_s L^s)X_t + u_t$$

$$Y_t = \mu + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_r Y_{t-r} + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_s X_{t-s} + u_t$$

Si tinguéssim més d'una variable explicativa, el model seria encara més general; X_j $\forall\ j = 0, 1, \ldots, k$, que dóna lloc a $AD(r;\ s_1, s_2, ..., s_k)$. Al contrari, models més simples són el $RD(s)$, que podria definir-se com un $AD(0, s)$, i el $AR(r)$, definit com un $AD(r, 0)$.

5) Altres models dinàmics. Normalment especifiquem models basats en la teoria econòmica però també podem fer-ho utilitzant la informació històrica disponible per decidir el nombre de retards associats de cada una de les variables inclosos en l'especificació. Entre els més habituals trobem:

a) Els **models ARIMA(p,d,q) univariants**:

Els models ARIMA(p,d,q) són models ARMA(p,q) que s'han hagut de diferenciar d vegades per a convertir-los en estacionaris pel que fa a la mitjana.

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p)Y_t = \mu + (\theta_0 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q)\varepsilon_t$$

$$\phi(L)Y_t = \mu + \theta(L)\varepsilon_t$$

b) Els **models de funció de transferència** es basen en la teoria de la cointegració, que combina l'econometria clàssica amb l'anàlisi de sèries temporals, que al seu torn tracta d'evitar la modelització de relacions espúries entre variables. L'anomenat model de mecanisme de correcció de l'error s'especifica de la manera següent:

La relació entre dues variables és espúria quan no és determinada per la relació entre aquestes variables sinó per altres causes, com per exemple una tercera variable.

$$Y_t = \mu + [B(L)/A(L)]X_t + [\phi(L)/\theta(L)]\varepsilon_t$$

1.1.2. Anàlisi i interpretació dels models dinàmics

A més d'interpretar els paràmetres estimats i la bondat de l'ajust, en els models s'han de tenir en compte altres aspectes relacionats amb els efectes que la temporalitat té en la interpretació econòmica dels paràmetres:

1) El model estimat ha de ser **estable**, és a dir, si es produeix una variació puntual de qualsevol X en un determinat moment t, la variable Y torna al seu valor d'equilibri i fa que l'efecte total derivat d'aquesta variació sigui finit. També es tracta d'un model estable si davant una variació permanent de X la variable Y evoluciona cap a un nou valor d'equilibri.

Si davant una variació puntual de X s'assoleix un nou valor d'equilibri de Y, es tracta d'un model **inestable explosiu**, però si no s'assoleix aquest nou valor d'equilibri el model és **inestable no explosiu**.

Vegem-ne un resum en la taula següent:

Model	Variació de X	Valor inicial X	V. transitori X	V. final X	V. inicial Y	V. transitori Y	V. final Y
Estable	Puntual	X_0	X_1	X_0	Y_0	Y_1	Y_0
Estable	Permanent	X_0	X_1	X_1	Y_0	Y_1	Y_1
Inestable expl.	Puntual	X_0	X_1	X_0	Y_0	Y_1	Y_1
Inestable no exp.	Puntual	X_0	X_1	X_0	Y_0	Y_1	$\nexists$
Inestable no exp.	Permanent	X_0	X_1	X_1	Y_0	Y_1	$\nexists$

L'anàlisi de l'estabilitat del model parteix de l'equació $A(L)Y_t = \mu + B(L)X_t + u_t$ i consisteix a comprovar que les arrels del polinomi autoregressiu $A(L)$ $(1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_r L^r = 0)$ caiguin fora del cercle de la unitat $(\forall L > 1)$. Si això es compleix, la Y recupera el seu valor d'equilibri davant una variació puntual de X, o assoleix un nou valor d'equilibri si la variació de X és permanent.

2) S'han de diferenciar els efectes contemporanis (**multiplicador d'impacte o contemporani**) dels no contemporanis que tenen les X sobre la Y (**multiplicadors de retardats o una vegada transcorreguts j períodes, i multiplicador total**).
Anem pre parts; donat el model:

Els multiplicadors són els paràmetres del model i es defineixen com a $\beta_j = \frac{\delta Y}{\delta X_j}$ $j = 2, \ldots, k$, és a dir, ens expliquen el canvi a Y davant un canvi unitari de qualsevol X.

$$Y_t = \mu + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_r Y_{t-r} + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_s X_{t-s} + u_t$$

definim els conceptes de multiplicador següents:

a) El **multiplicador d'impacte o contemporani** (m_0) és la variació de Y_t davant una variació unitària de X_t, és a dir, $m_0 = \dfrac{\partial Y_t}{\partial X_t} = \beta_0$

b) L'**efecte multiplicatiu després d'haver transcorregut j períodes** (m_j) serà la variació produïda a Y_t davant una variació unitària de X_{t-j}, o sigui, $m_j = \dfrac{\partial Y_t}{\partial X_{t-j}}$.

Quan $m_j \neq \beta_j$, observem l'existència d'una dependència implícita de les variables endògenes retardades.

c) Finalment, el **multiplicador total** (m_T) és la suma de tots els multiplicadors contemporanis i no contemporanis, és a dir, $m_T = \sum_{j=0}^{\infty} m_j$.

Per a calcular els multiplicadors és molt útil utilitzar l'expressió següent:

$$Y_t = \mu/A(L) + [B(L)/A(L)]X_t + u_t/A(L) = \mu' + D(L)X_t + v_t$$

Quan es tracta d'un model RD trobar els multiplicadors és relativament senzill, ja que es redueixen a la identificació de les diferents β que componen el polinomi $B(L)$. En canvi, quan existeix part AR s'ha de calcular $D(L) = \frac{B(L)}{A(L)}$. A més a més, el m_0 i el m_T seran respectivament els valors de $D(L)$ quan L és zero i u, respectivament.

donde $D(L) = \delta_0 + \delta_1 L + \delta_2 L^2 + \ldots$ ja que:

$$m_0 = \frac{\partial Y_t}{\partial X_t} = D(0) = \delta_0$$

$$m_j = \frac{\partial Y_t}{\partial X_{t-j}} = \delta_j$$

$$m_T = \sum_{j=0}^{\infty} m_j = D(1) = \delta_0 + \delta_1 + \delta_2 + \ldots$$

3) El **retard mitjà** i el **retard medià** també ens ajuden a interpretar els resultats de l'estimació dels models dinàmics.

a) El retard medià no és més que la mitjana ponderada dels coeficients del polinomi $D(L)$ del model $Y_t = \mu' + D(L)X_t + v_t$, és a dir:

$$Retard\,mit\,j = \frac{\sum_{j=0}^{\infty} j \cdot \delta_j}{\sum_{j=0}^{\infty} \delta_j} = \frac{D'(1)}{D(1)} = \frac{B'(1)}{B(1)} - \frac{A'(1)}{A(1)}$$

on $D'(L)$, $B'(L)$ i $A'(L)$ són les derivades de $D(L)$, $B(L)$ i $A(L)$, respecte a L i tots aquests polinomis valorats a $L = 1$. De manera que un valor elevat del retard mitjà implica que la contribució al comportament de Y dels períodes allunyats en el temps és alta, cosa que indica la concentració o dilució dels efectes de les variables exògenes.

b) El retard medià indica el moment en què s'assoleix el 50% de la variació total que es produeix a Y a causa d'una variació a X.

També es poden fer aquests mateixos càlculs amb un full de càlcul com Excel.

Vegem-ne un exemple utilitzant la consola de R com si fos una calculadora. Si volguéssim comptar el nombre de cerveses venudes trimestralment d'una determinada marca (C) en funció del preu mitjà de la cervesa (P) i la despesa trimestral en campanyes de publicitat que ha fet aquesta marca de cervesa (CP) i, si fos necessari, les

vendes fetes d'aquesta mateixa marca en trimestres anteriors, per a un trimestre t determinat, podríem especificar qualsevol dels següents models:

1) Model 1. Som partidaris de la simplicitat i realment expliquem les nostres vendes en funció del preu i de les despeses que s'han generat en la campanya publicitària del període:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 CP_t + u_t \qquad t = 1, \ldots, T.$$

Aquest model potser no és gaire bo, fàcilment hi podríem detectar una autocorrelació dels residus derivada d'una mala especificació o d'una omissió de variables rellevants, ja que no hem considerat cap relació dinàmica en el nostre model.

2) Model 2. Estem convençuts que les despeses en les campanyes publicitàries anteriors tenen molt d'impacte en les vendes dels trimestres posteriors senzillament perquè no és tan fàcil que els nostres clients s'oblidin de les cançons encomanadisses que conscientment hem seleccionat per a les campanyes de primavera i estiu; el nostre model passaria a ser:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 CP_t + \beta_3 CP_{t-1} + \beta_4 CP_{t-2} + \beta_5 CP_{t-3} + u_t \qquad t = 1, \ldots, T.$$

3) Model 3. Considerem que són les vendes dels trimestres anteriors sobre les quals es basen les vendes actuals:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 CP_t + \alpha_1 C_{t-1} + \alpha_2 C_{t-2} + u_t \qquad t = 1, \ldots, T.$$

4) Model 4. Estem d'acord amb les dues teories anteriors:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 CP_t + \beta_3 CP_{t-1} + \beta_4 CP_{t-2} + \beta_5 CP_{t-3} + \alpha_1 C_{t-1} + \alpha_2 C_{t-2} + u_t$$

$$t = 1, \ldots, T.$$

Una vegada triada l'especificació que més s'ajusti al nostre entorn, hauríem de valorar aquest model utilitzant la informació que tenim. Suposem que tenim informació trimestral dels últims 10 anys. És clar que en aquest exemple tractem amb dades de periodicitat trimestral que implica un conjunt de 40 dades; és a dir, que en el nostre cas, quan especifiquem $t = 1, \ldots, T$, la nostra $T = 40$. Suposem ara que les estimacions dels models proposats són:

1) Model 1:

$$C_t = 73.96 + 16.83 P_t + 0.07 CP_t$$

2) Model 2:

$$C_t = 59.52 + 20.51P_t + 0.10CP_t + 0.02CP_{t-1} - 0.05CP_{t-2} + 0.09CP_{t-3}$$

3) Model 3:

$$C_t = 127.62 + 22.44P_t + 0.35CP_t + 0.55C_{t-1} + 0.15C_{t-2}$$

4) Model 4:

$$C_t = 105.16 + 25.61P_t + 0.5CP_t + 0.35CP_{t-1} - 0.15CP_{t-2} + 0.12CP_{t-3} + 0.65C_{t-1} + 0.25C_{t-2}$$

Primer determinarem el tipus i l'estabilitat dels models estimats:

1) Model 1. Es tracta d'un model estàtic. No té sentit parlar d'estabilitat.

2) Model 2. És un model dinàmic $RD(3)$. La part de retards distribuïts d'ordre 3 associada a CP serà $(0.10 + 0.02L - 0.05L^2 + 0.09L^3)$CPt. Per definició tots els models de retards distribuïts són sempre estables.

3) Model 3. És un model dinàmic $AR(2)$. La part autoregressiva d'ordre 2 serà $(1 - 0.55L - 0.15L^2)$. Per analitzar l'estabilitat del model primer el reescrivim:

$$(1 - 0.55L - 0.15L^2)\hat{C}_t = 127.62 + 22.44P_t + 0.35CP_t$$

després, calculem les arrels del polinomi de retards i mirem si cauen o no fora del cercle de la unitat. Efectuant els càlculs manualment hem d'escriure en la finestra d'instruccions:

```
m3 <- c(1,-0.55,-0.15)
polyroot(m3)
```

En seleccionar el que s'ha escrit i prémer *Executar*, obtindrem en la finestra de resultats:

```
> m3 <- c(1,-0.55,-0.15)
> polyroot(m3)
[1]  1.333333+0i -5.000000-0i
```

Així, es tracta d'un model estable perquè les seves arrels són més grans que la unitat en valor absolut; $L_i > |1|$

4) Model 4. És un model dinàmic $AD(2,3)$. La part autoregressiva d'ordre 2 serà $(1 - 0.65L - 0.25L^2)$, i la part de retards distribuïts (d'ordre 3) associada a CP serà $(0.5 + 0.35L - 0.15L^2 + 0.12L^3)$CPt. Una vegada reescrit el model:

$$(1 - 0.65L - 0.25L^2)\hat{C}_t = 105.16 + 25.61P_t + (0.5 + 0.35L - 0.15L^2 + 0.12L^3)CP_t$$

n'analitzem l'estabilitat de la mateixa manera que hem fet abans. En la finestra de resultats apareix:

```
> m4 <- c(1,-0.65,-0.25)
> polyroot(m4)
[1]  1.085372+0i -3.685372+0i
```

Ara calcularem els multiplicadors amb els models expressats en funció dels seus polinomis de retards:

1) Model 1. No és procedent.

2) Model 2. És un model estable $RD(3)$ que es pot expressar de la manera següent:

$$\hat{C}_t = 59.52 + 20.51P_t + (0.10 + 0.02L - 0.05L^2 + 0.09L^3)CP_t$$

Sabem que en aquest tipus de models els paràmetres poden interpretar-se com els multiplicadors contemporanis o retardats de CP sobre C. A més també podem calcular-ne el retard mitjà i el retard medià de la manera següent:

Concepte	Càlcul	Resultat
Multiplicador contemporani	β_2	0.10
Multiplicador retardat 1 període	β_3	0.02
Multiplicador retardat 2 períodes	β_4	-0.05
Multiplicador retardat 3 períodes	β_5	0.09
Multiplicador total	$\beta_2 + \beta_3 + \beta_4 + \beta_5$	$0.10 + 0.02 - 0.05 + 0.09 = 0.16$
Retard mitjà	$\frac{B'(1)}{B(1)}$	$\frac{0.02-2*0.05+3*0.09}{0.16} = \frac{0.19}{0.16} = 1.1875$
Retard medià	t en assolir 0.08	9 mesos i 18 dies del mateix any

3) Model 3. És un model estable $AR(2)$ que es pot expressar de la manera següent:

$$(1 - 0.55L - 0.15L^2)\hat{C}_t = 127.62 + 22.44P_t + 0.35CP_t$$

Els seus multiplicadors contemporanis o retardats, els retards mitjans i medians de CP sobre C són:

Concepte	Càlcul	Resultat
Multiplicador contemporani	β_2	0.35
Multiplicador retardat 1 període	$\alpha_1 * \beta_2$	0.19
Multiplicador retardat 2 períodes	$(\alpha_1^2 + \alpha_2) * \beta_2$	0.16
Multiplicador retardat 3 períodes	$(\alpha_1^3 + 2\alpha_1\alpha_2) * \beta_2$	0.12
Multiplicador total	$\delta_0 + \delta_1 + \delta_2 + \delta_3$	$0.35 + 0.19 + 0.16 + 0.12 = 0.82$
Retard mitjà	$\frac{B'(1)}{B(1)} - \frac{A'(1)}{A(1)}$	$\frac{0.00}{0.35} - \frac{-0.55-2*0.15}{1-0.55-0.15} = 0 - \frac{-0.85}{0.30} = 2.83$
Retard medià	t en assolir 0.41	3 mesos i 18 dies del pròxim any

4) Model 4. És un model estable $AD(2,3)$ que es pot expressar de la manera següent:

$$(1 - 0.65L - 0.25L^2)\hat{C}_t = 105.16 + 25.61P_t + (0.5 + 0.35L - 0.15L^2 + 0.12L^3)CP_t$$

Els seus multiplicadors contemporanis o retardats, els retards mitjans i medians de CP sobre C són:

Concepte	Càlcul	Resultat
Multiplicador contemporani	β_2	0.50
Multiplicador retardat 1 període	$\beta_3 + \alpha_1 * \beta_2$	0.675
Multiplicador retardat 2 períodes	$\beta_4 + \alpha_1 * \beta_3 + (\alpha_1^2 + \alpha_2) * \beta_2$	0.414
Multiplicador retardat 3 períodes	$\beta_5 + \alpha_1 * \beta_4 + (\alpha_1^2 + \alpha_2) * \beta_3 + (\alpha_1^3 + 2\alpha_1\alpha_2) * \beta_2$	0.558
Multiplicador total	$\delta_0 + \delta_1 + \delta_2 + \delta_3$	$0.5 + 0.675 + 0.414 + 0.558 = 2.146$
Retard mitjà	$\frac{B'(1)}{B(1)} - \frac{A'(1)}{A(1)}$	$\frac{0.35-2*0.15+3*0.12}{0.5+0.35-0.15+0.12} - \frac{-0.65-2*0.25}{1-0.65-0.25} = \frac{0.41}{0.82} - \frac{-1.15}{0.10} = 12$
Retard medià	t en assolir 0.85	10 mesos i 6 dies del pròxim any

En resum:

Model	m_0	m_1	m_2	m_3	m_T	Retard mitjà	Retard medià
Model 1	∄	∄	∄	∄	∄	∄	∄
Model 2	0.10	0.02	−0.05	0.09	0.16	1.1875	9 mesos i 18 dies del mateix any
Model 3	0.35	0.19	0.16	0.12	0.82	2.83	3 mesos i 18 dies del pròxim any
Model 4	0.50	0.675	0.414	0.558	2.146	12	10 mesos i 6 dies del pròxim any

1.1.3. Mètodes d'estimació

1) Estimació per MCO en models amb variables exògenes retardades. Són els models de retards distribuïts $RD(s)$. Si el terme de pertorbació compleix totes les hipòtesis bàsiques (té una matriu de variàncies i covariàncies escalar), podem utilitzar els estimadors MCO sense problemes perquè seran **centrats, eficients i consistents**. No obstant això, poden aparèixer alguns problemes:

a) A mesura que augmenta el nombre de retards de la variable exògena tindrem menys graus de llibertat i disminuirà la fiabilitat de l'estimació.

b) Es pot presentar un nivell de multicol·linealitat elevada per a utilitzar com a regressors la mateixa variable referida a diferents moments del temps.

Recordeu que el mòdul 2 d'aquest manual tracta de la multicol·linealitat.

En cas que el terme de pertorbació fos no esfèrico esfèric, hauríem d'utilitzar MCG per a obtenir estimadors eficients.

Podeu consultar el mòdul 2 d'aquest manual per aprofundir en l'estimació per MCG.

Finalment, si ens trobéssim amb l'estimació d'un model de retards distribuïts amb un nombre infinit de retards no podríem estimar per falta de dades i hauríem de transformar el model utilitzant la hipòtesi de Koyck.

2) Estimació per MCO en models amb variables explicatives incorrelacionades amb el terme de pertorbació. Són models dinàmics en què utilitzem com a variables explicatives retards de la mateixa variable endògena, i per tant, podem tenir problemes d'estimació perquè tindrem regressors estocàstics. La clau està a veure si aquestes estan o no correlacionades amb el terme de pertorbació.

Perquè el terme de pertorbació (u_t) estigui incorrelacionat amb els regressors la matriu de variàncies i covariàncies ha de ser escalar. D'aquesta manera, totes les variables explicatives tindran una distribució independent del terme de pertorbació. El que passa és que Y_{t-1} no és independent del terme de pertorbació a t ? 1 i en períodes anteriors, pero sí que és independent a t i en períodes posteriors. Així doncs, utilitzant el teorema de Mann-Wald, arribem a la conclusió que l'estimador MCO és **asimptòticament centrat, consistent i**, si U es distribueix segons una normal, també **asimptòticament eficient**.

3) Estimació per MCO en models amb variables explicatives correlacionades amb el terme de pertorbació. En aquests models dinàmics també s'utilitzen com a regressors variables endògenes retardades i estan correlacionades amb el terme de pertorbació. Per tant, les estimacions per MCO són **centrades** (i el biaix no tendeix a zero en augmentar la grandària de la mostra), **inconsistents** perquè no es compleix el teorema de Mann-Wald **i ineficients** per la no-esfericitat del terme de pertorbació u_t.

4) Mètodes d'estimació alternatius:

a) El mètode d'estimació de *variables instrumentals (VI)* assegura la consistència dels estimadors i s'utilitza tant en models dinàmics amb endògenes retardades correlacionades amb un terme de pertorbació autocorrelacionat com en models en què hi ha correlació entre variables explicatives i el terme de pertorbació perquè no es compleix el supòsit d'exogenitat dels regressors.

Recordeu que el teorema de Mann-Wald assegura la consistència dels estimadors.

En el cas que ens afecta partim d'un model $Y = ZB + U$, on Z és la matriu de dimensió $T \times k$ de variables explicatives del model dinàmic (que inclou variables endògenes retardades). Aplicant MCO els estimadors eren inconsistents perquè hi havia correlació entre Z_t i u_t; per tant $E[Z_t' u_t] \neq 0$ i no podíem aplicar el teorema de Mann-Wald.

El que es proposa amb el mètode de les VI és definir una nova matriu W_t, amb tantes variables com Z_t, de manera que aquestes noves variables (instruments) estiguin tan correlacionades com sigui possible amb les variables inicials i no ho estiguin amb el terme de pertorbació del model per poder així aplicar el teorema de Mann-Wald i estimar per MCO.

El problema radica llavors a trobar els instruments adequats per a fer l'estimació. Podem trobar-nos davant tres possibilitats:

1) Obtenir una *proxy* que ens serveixi d'instrument per a definir la nova matriu W_t.
2) Utilitzar com a instrument la variable que millor expliqui el comportament de la variable endògena i així poder definir la nova matriu W_t.
3) Obtenir una estimació de l'instrument a partir d'una regressió auxiliar per a definir la nova matriu W_t. Quan la variable que ens crea problemes és l'endògena retardada, aquesta regressió auxiliar es calcula utilitzant com a endògena l'endògena retardada i com a explicatives totes les variables explicatives (excepte la mateixa endògena retardada) del model inicial retardades un període. Si utilitzem aquesta última possibilitat, els estimadors obtinguts es denominen **estimadors per mínims quadrats en dues etapes** i tenen la particularitat de proporcionar unes estimacions més eficients dels paràmetres inicials que si utilitzéssim un altre tipus d'instruments per a l'endògena retardada.

b) El mètode d'estimació de *mínims quadrats no lineals (MCNL)*, basat a minimitzar la suma dels quadrats dels residus.

c) El mètode d'estimació de *màxima versemblança (MV)* consisteix a maximitzar el logaritme neperià de la funció de versemblança.

Els estimadors d'aquests dos últims mètodes coincideixen si el terme de pertorbació segueix una llei de distribució normal.

Fixeu-vos que l'estimador de mínims quadrats en dues etapes està més que justificat en el model especificat perquè inclou la variable endògena retardada un període entre els seus regressors.

Vegem ara com fer amb R-Commander una estimació per VI utilitzant mínims quadrats en dues etapes. Per a això recuperarem l'essència de l'exemple. Intentarem explicar el nombre de cerveses venudes trimestralment d'una determinada marca (C) en funció de la variació del preu mitjà de la cervesa (P) en el trimestre actual i en l'anterior; la despesa trimestral en campanyes de publicitat que ha fet en els tres últims trimestres (CP), el percentatge de població adulta que pot consumir cervesa (A). A més a més, considerarem que les vendes fetes d'aquesta mateixa marca en el trimestre anterior també és determinant per a calcular les vendes actuals. Així, el nostre nou model serà:

$$C_t = \mu + \beta_0 CP_t + \beta_1 CP_{t-1} + \beta_2 CP_{t-2} + \alpha_0 P_t + \alpha_1 P_{t-1} + \delta_0 A_t + \gamma_0 C_{t-1} + u_t$$

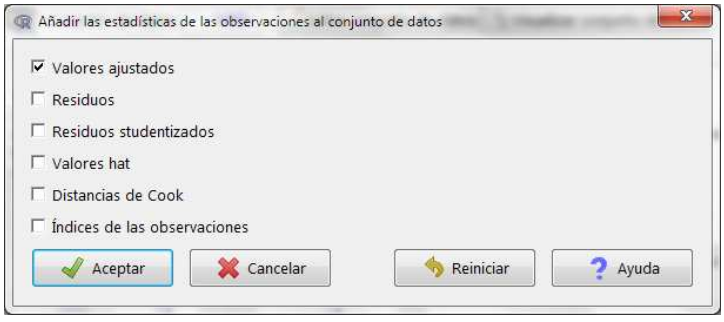
$$t = 1, \ldots, T.$$

Per estimar el model seguirem els passos següents:

a) Llegir la base de dades denominant-la MRD i ajustar el model per MCO excloent la variable C_{t-1}, denominant-lo *Modelo*1. En la finestra de resultats ens apareix el següent:

```
> Modelo1 <- lm(C~A+CP+CP_1+CP_2+P+P_1, data=MRD)
> summary(Modelo1)
Call:
lm(formula = C ~ A + CP + CP_1 + CP_2 + P + P_1, data = MRD)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07805 -0.03228 -0.01139  0.02653  0.11847

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) -4.81843    0.36083 -13.354 2.13e-14 ***
A            0.18891    0.29194   0.647 0.522339
CP           1.40129    0.36461   3.843 0.000564 ***
CP_1         0.72265    0.50591   1.428 0.163171
CP_2        -0.91264    0.35998  -2.535 0.016498 *
P            0.26388    0.04881   5.406 6.71e-06 ***
P_1         -0.15628    0.05605  -2.788 0.008973 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05319 on 31 degrees of freedom
Multiple R-squared:  0.9975,  Adjusted R-squared:  0.9971
F-statistic:  2090 on 6 and 31 DF,  p-value: < 2.2e-16
```

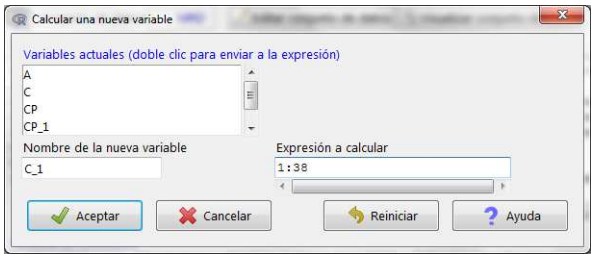
b) Guardar els valors ajustats afegint una nova variable a la base de dades denominada *fitted.Modelo*1. Per a això seguirem la ruta *Models / Afegir les estadístiques de les observacions a les dades...*, i en el quadre de diàleg que aparegui marcarem únicament l'opció de valors ajustats tal com es mostra a continuació:



c) Construir la variable C_{t-1}. Això ho podem fer directament en la base de dades: en primer lloc, utilitzem la funció `nrow` per a assegurar-nos del nombre exacte de les observacions de la nostra base de dades, que denominarem *n*.

```
n <- nrow (MRD)
```

En segon lloc, crearem una nova variable anomenada C_1, utilitzant l'opció *Dades / Modificar variables del conjunt de dades actiu / Calcular una nova variable* que vagi des d'1 fins a *n*:



Una vegada creada la variable, que podia haver estat també una columna de zeros, d'uns o de valors perduts, simplement reemplaçant l'1 : 38 per 0, 1 o *NA*, respectivament, reemplaçarem el de la variable C_1 pel de la variable C a partir de la segona observació i deixant buida la primera fila. Això és molt senzill utilitzant el següent codi:

Per a més informació sobre el que fa aquest codi podeu consultar el mòdul 1 del manual *Matemáticas y Estadística con R.*

```
for (i in 1:37) {MRD$C_1[i+1] <- MRD$C[i]}
MRD$C_1[1] <- NA
```

Podem comprovar que ho hem fet bé visualitzant el conjunt de dades:

MRD

	C	CP	CP_1	CP_2	P	P_1	A	fitted.Modelo1	C_1
1	4.634729	7.796058	7.725330	7.654917	-0.6931472	-0.6931472	0.09139872	4.645290	NA
2	4.699257	7.850493	7.796058	7.725330	-0.6931472	-0.6931472	0.16778523	4.722850	4.634729
3	4.763786	7.881560	7.850493	7.796058	-0.6931472	-0.6931472	0.13737600	4.735427	4.699257
4	4.828314	7.967280	7.881560	7.850493	-0.3566750	-0.6931472	0.08050962	4.906362	4.763786
5	4.940471	8.024535	7.967280	7.881560	-0.3566750	-0.3566750	0.16395357	4.983363	4.828314
6	5.052629	8.130059	8.024535	7.967280	-0.3566750	-0.3566750	0.18042098	5.097487	4.940471
7	5.164786	8.231376	8.130059	8.024535	-0.5108256	-0.3566750	0.13638264	5.214468	5.052629
8	5.335131	8.313852	8.231376	8.130059	-0.5108256	-0.5108256	0.14750044	5.333144	5.164786
9	5.480639	8.368925	8.313852	8.231376	-0.3566750	-0.5108256	0.18322665	5.424878	5.335131
10	5.545177	8.439232	8.368925	8.313852	0.0953102	-0.3566750	0.11559507	5.570327	5.480639
11	5.645447	8.551595	8.439232	8.368925	-0.2231435	0.0953102	0.17827916	5.585495	5.545177
12	5.849325	8.619027	8.551595	8.439232	0.0000000	-0.2231435	0.08542789	5.788134	5.645447
13	6.075346	8.712595	8.619027	8.551595	0.6931472	0.0000000	0.18180602	6.031674	5.849325
14	6.146329	8.783243	8.712595	8.619027	0.7419373	0.6931472	0.13118231	6.031729	6.075346
15	6.240276	8.841304	8.783243	8.712595	0.9162907	0.7419373	0.15595625	6.121811	6.146329
16	6.326149	8.949625	8.841304	8.783243	1.0296194	0.9162907	0.11122376	6.245288	6.240276
17	6.393591	9.049937	8.949625	8.841304	0.9555114	1.0296194	0.14665164	6.380569	6.326149
18	6.489205	9.152605	9.049937	8.949625	0.9932518	0.9555114	0.16918275	6.523866	6.393591
19	6.622736	9.286190	9.152605	9.049937	0.9555114	0.9932518	0.19783198	6.683256	6.489205
20	6.673298	9.296885	9.286190	9.152605	1.0986123	0.9555114	0.14955891	6.735619	6.622736
21	6.744059	9.366575	9.296885	9.286190	1.5040774	1.0986123	0.08129425	6.790822	6.673298
22	6.814543	9.438113	9.366575	9.296885	1.3609766	1.5040774	0.12606960	6.838997	6.744059
23	6.892642	9.496947	9.438113	9.366575	1.2809338	1.3609766	0.08754240	6.903501	6.814543
24	6.962243	9.533872	9.496947	9.438113	1.3083328	1.2809338	0.12945782	6.960128	6.892642
25	7.006695	9.610525	9.533872	9.496947	1.3083328	1.3083328	0.08751287	7.028325	6.962243
26	7.057898	9.673068	9.610525	9.533872	1.1939225	1.3083328	0.10901158	7.111529	7.006695
27	7.109879	9.716616	9.673068	9.610525	1.0296194	1.1939225	0.16130477	7.132196	7.057898
28	7.170120	9.727585	9.716616	9.673068	1.2237755	1.0296194	0.13504321	7.193910	7.109879
29	7.229114	9.750977	9.727585	9.716616	1.5260563	1.2237755	0.12738576	7.242847	7.170120
30	7.275865	9.785998	9.750977	9.727585	1.6292405	1.5260563	0.09479042	7.272643	7.229114
31	7.331715	9.838576	9.785998	9.750977	1.6486586	1.6292405	0.13229432	7.346362	7.275865
32	7.386471	9.888425	9.838576	9.785998	1.5040774	1.6486586	0.11563319	7.377914	7.331715
33	7.483244	9.979012	9.888425	9.838576	1.3862944	1.5040774	0.17261532	7.495172	7.386471
34	7.533694	10.010547	9.979012	9.888425	1.4350845	1.3862944	0.18998090	7.593892	7.483244
35	7.591862	10.056809	10.010547	9.979012	1.4109870	1.4350845	0.11578445	7.570833	7.533694
36	7.639642	10.091957	10.056809	10.010547	1.3609766	1.4109870	0.15743518	7.623175	7.591862
37	7.707063	10.148393	10.091957	10.056809	1.2527630	1.3609766	0.15052175	7.663390	7.639642
38	7.762171	10.183881	10.148393	10.091957	1.1314021	1.2527630	0.10899437	7.698868	7.707063

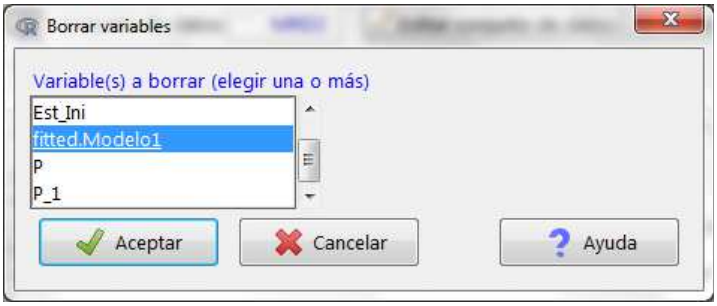
d) Construir la matriu W creant una nova base de dades idèntica a l'anterior, que anomenarem MRD2 mitjançant la instrucció:

```
MRD2 <- MRD
```

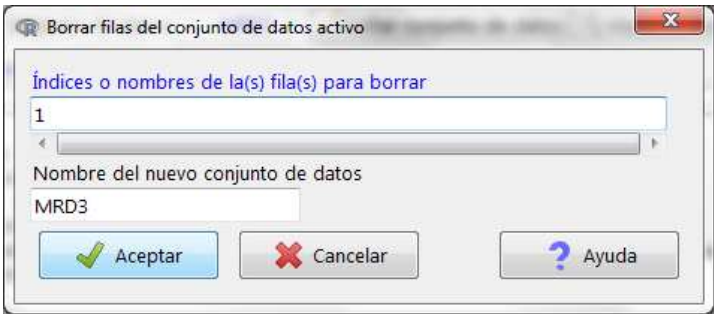
En aquesta nova base de dades inserim la variable d'estimació del model retardada un període i l'anomenarem Est_Ini; això ho farem com abans, quan hem creat C_1. És a dir:

```
MRD2$Est_Ini <- 1:38
for (i in 1:37) {MRD2$Est_Ini[i+1] <- MRD2$fitted.Modelo1[i]}
MRD2$Est_Ini[1] <- NA
```

Després eliminarem de la base de dades la variable $fitted.Modelo1$ mitjançant la ruta *Dades / Modificar variables del conjunt de dades actiu / Eliminar variables del conjunt de dades*, i triant la variable que hem d'eliminar en el quadre de diàleg:



A continuació també eliminarem la primera observació utilitzant el menú *Dades / Conjunt de dades actiu / Esborrar fila(es) del conjunt de dades actiu*. Com veiem, aquí tenim la possibilitat d'anomenar novament la base de dades, fem que sigui MRD3:



En aquests moments seguirem tenint el mateix nombre de variables però ara tindrem una observació menys. A continuació crearem una nova variable anomenada $UNOS$ que contingui una columna d'uns utilitzant de nou el menú *Dades / Modificar variables del conjunt de dades actiu / Calcular una nova variable*.

Finalment, abocarem aquesta nova base dades en l'espai de treball per crear les matrius que necessitem mitjançant la funció `attach`. La matriu W contindrà les variables: $UNOS$, CP, CP_1, CP_2, P, P_1, A, Est_Ini i en l'ordre indicat:

```
attach (MRD3)
W <- matrix (c(UNOS, CP, CP_1, CP_2, P, P_1, A, Est_Ini),37,8)
```

Recordeu que el mòdul 2 del manual *Matemáticas y Estadística con R* tracta en profunditat les operacions amb matrius.

e) Construir la matriu Z. Partirem de la base MRD3 que acabem de crear però ara inclourem les variables $UNOS$, CP, CP_1, CP_2, P, P_1, A, C_1:

```
Z <- matrix (c(UNOS, CP, CP_1, CP_2, P, P_1, A, C_1),37,8)
```

f) Construir la matriu Y quedant-nos únicament amb la variable endògena C.

```
Y <- matrix (c(C),37,1)
```

g) Calcular la matriu $W'Z$. Primer calcularem la matriu transposada amb la funció `t` i després multipliquem mitjançant l'operador $\%*\%$.

També ho podíem haver fet directament mitjançant el codi $tWZ <- t(W)\%*\%Z$.

```
tW <- t(W)
tWZ <- tW%*%Z
```

h) Calcular la matriu $(W'Z)^{-1}$. Per a calcular la matriu inversa utilitzarem la funció `solve`.

Molt més curt hauria estat fent $ItWZ <- solve(t(W)\%*\%Z)$.

```
ItWZ <- solve(tWZ)
```

i) Calcular la matriz $(W'Y)$.

```
tWY <- t(W)%*%Y
```

j) I finalment obtenir l'estimació dels paràmetres calculant la matriu: $(W'Z)^{-1}(W'Y)$. Això és:

D'aquesta manera ens hauríem estalviat un munt de passos: $B <- solve(t(W)\%*\%Z, t(W)\%*\%Y)$.

```
B <- ItWZ%*%tWY
```

Amb això l'estimació per mínims quadrats en dues etapes del nostre model dinàmic és:

$$\hat{C}_t = -1.15 + 0.83CP_t - 0.02CP_{t-1} - 0.58CP_{t-2} + 0.11P_t - 0.14P_{t-1} + 0.12A_t + 0.84C_{t-1}$$

1.2. Models de regressió multiequacionals

És evident que no tots els models economètrics pretenen explicar el comportament d'una única variable endògena mitjançant altres variables explicatives predeterminades (exògenes o endògenes retardades). A vegades, per a explicar una variable és necessari considerar altres variables endògenes com a explicatives, de manera que s'obté més d'una equació que s'ha d'estimar.

Considerarem que una variable és endògena si no está controlada per l'analista i s'ha d'explicar pel model. Mentre que una variable serà exògena si se'n pot predeterminar el valor i, per tant, no s'explica pel model.

1.2.1. Hipòtesis bàsiques i formulació general d'un model multiequacional

En un principi en un model multiequacional s'han de complir les següents hipòtesis:

Recordem que en forma matricial el conjunt de variables endògenes es representa pel vector Y i el conjunt de variables predeterminades per la matriu Z.

1) Disposar de G variables endògenes que ha d'explicar el nostre model. Això significa especificar G equacions perquè el model sigui complet. Cada variable endògena serà explicada per k variables predeterminades (entre exògenes i endògenes retardades). No és obligatori que en cada equació figurin totes les variables endògenes i predeterminades però sí que la relació existent sigui lineal. Per tant, un model multiequacional podria ser:

$$\begin{aligned}
&\beta_{11}Y_{1t} + \beta_{12}Y_{2t} + \ldots + \beta_{1G}Y_{Gt} + \gamma_{11}Z_{1t} + \ldots + \gamma_{1k}Z_{kt} = u_{1t},\\
&\beta_{21}Y_{1t} + \beta_{22}Y_{2t} + \ldots + \beta_{2G}Y_{Gt} + \gamma_{21}Z_{1t} + \ldots + \gamma_{2k}Z_{kt} = u_{2t},\\
&\ldots \qquad\qquad \ldots \qquad\qquad \ldots \qquad\qquad \ldots \qquad\qquad ,\\
&\beta_{G1}Y_{1t} + \beta_{G2}Y_{2t} + \ldots + \beta_{GG}Y_{Gt} + \gamma_{G1}Z_{1t} + \ldots + \gamma_{Gk}Z_{kt} = u_{Gt},
\end{aligned}$$

on $t = 1, 2, \ldots, T$ observacions. A més a més s'ha de suposar que $\beta ii = 1$.

2) Encara que els termes de pertorbació de cada equació tindran mitja nul·la i seran esfèrics, suposarem que hi ha termes de pertorbació referits a diferents equacions del model que puguin presentar autocorrelació contemporània.

Una vegada definit un model multiequacional, vegem les diferents maneres d'escriure'l:

Matricialment, això s'escriu $E[U] = 0$ i $E[UU'] = \Sigma \otimes I$, on $\otimes$ és el producte de Kronecker que implica que cada element de la primera matriu es multiplica per la segona, i U és la matriu de dimensió $TG \times TG$ formada per totes les T observacions de cada un dels G termes de pertorbació.

1) La **forma estructural** consisteix a escriure en cada una de les G equacions la combinació lineal de totes les variables endògenes i predeterminades, i igualar-les als seus termes de pertorbació.

2) La **forma reduïda** és quan expressem cada variable endògena en funció de les variables predeterminades.

3) La **forma final** és aquella en què les variables endògenes estan en funció únicament de variables exògenes i exògenes retardades. El punt de partida per a obtenir aquesta formulació a partir de la forma estructural requeriria utilitzar l'operador de retards.

Un exemple de les diferents maneres d'escriure un model multiequacional es troba en el manual de l'assignatura d'*Econometria*.

1.2.2. Tipologia de models multiequacionals

Es distingeixen quatre tipus de models multiequacionals:

1) Els **models multiequacionals no relacionats** són aquells en què un conjunt de variables predeterminades expliquen el comportament de G variables endògenes sense que això suposi una relació entre elles, és a dir, que no n'hi hagi una que expliqui l'altra ni tampoc existeixi correlació entre els termes de pertorbació de les diferents equacions.

2) Els **models multiequacionals aparentment no relacionats** són aquells en què també un conjunt de variables predeterminades expliquen el comportament de G variables endògenes. No obstant això, en aquest cas, encara que no hi hagi relació directa entre variables endògenes (una explica l'altra), sí que es posa de manifest una relació indirecta, ja que hi ha correlació entre els termes de pertorbació de les diferents equacions.

3) Els **models multiequacionals recursius** són aquells en què un conjunt de variables predeterminades expliquen el comportament de G variables endògenes i, a més, es produeix una relació de causalitat unidireccional directa entre les variables endògenes. No obstant això, no hi ha correlació entre els termes de pertorbació de les diferents equacions.

4) Finalment, els **models multiequacionals integrats** constitueixen el cas general, en què no es dóna cap dels casos particulars que hem esmentat abans. Hi haurà relacions directes entre les variables endògenes i, a més, correlació entre els termes de pertorbació de les diferents equacions.

1.2.3. El problema de la identificació

Recordeu que s'ha d'estudiar la identificació per a cada equació individualment. Un model estarà no identificat si hi ha alguna equació que no ho estigui. Si el model està identificat, per l'únic fet que hi hagi una equació sobreidentificada, el model estarà sobreidentificat. Així doncs, un model estarà exactament identificat quan totes les equacions ho estiguin.

Per poder estimar un model multiequacional primer hem de tractar la identificació d'aquest, ja que es pot donar el cas que no disposem de prou informació per a poder resoldre el sistema d'equacions i, per tant, valorar els paràmetres del model. Així, ens podem trobar davant tres tipus de situacions:

1) Models amb **equacions no identificades**, és a dir, quan no tenim prou informació per a estimar els paràmetres de la forma estructural de l'equació. Estaríem davant un sistema d'equacions incompatible. Aquesta equació s'haurà de reespecificar mitjançant la incorporació de nova informació en aquesta. El més habitual és incorporar

restriccions lineals sobre els paràmetres de les variables endògenes o predeterminades.

2) Models amb **equacions sobreidentificades**, això és, en què hi ha més d'una combinació de valors estimats possible dels paràmetres estructurals que complirien la relació entre variables registrada en l'equació. Es corresponen amb sistemes d'equacions compatibles indeterminats.

Alguns exemples de restriccions són $\beta_{ij} + \beta_{ih} = 0$ o $\beta_{ij} = 0$.

3) Models amb **equacions exactament identificades** i que, per tant, a partir de les variables incloses en el model, solament podem obtenir una única estimació dels paràmetres estructurals. Es corresponen amb sistemes d'equacions compatibles determinats.

Per les característiques dels diferents models multiequacionals, solament ens hem de centrar en l'estudi de la identificació dels models d'equacions simultànies integrats. I el problema de la identificació se centra en la forma estructural. Aquest problema es resol mitjançant les denominades condicions de rang i d'ordre:

No oblideu que els models multiequacionals no relacionats, els aparentment no relacionats i els recursius sempre estan identificats. A més, la forma reduïda d'un model també està sempre identificada.

1) La **condició de rang** consisteix a calcular el rang d'una matriu $(A\phi)$, de manera que aquest és $G-1$, l'equació està identificada i no ho estarà altrament. $A = (B|\Gamma)$ és la matriu de dimensió $G \times (G+k)$, formada per tots els paràmetres de la forma estructural del model. ϕ és la matriu de restriccions de dimensió $(G+k) \times q$, formada per tantes files com nombre de variables endògenes i predeterminades hi hagi, i tantes columnes (q) com restriccions presenti l'equació.

2) La **condició d'ordre** determina el tipus d'identificació. Si el nombre de restriccions és $G-1$, l'equació està exactament identificada i si és més gran que $G-1$, l'equació està sobreidentificada.

No oblideu que la condició d'ordre és una condició necessària, però no suficient. Podem tenir una equació no identificada que complís aquesta condició.

1.2.4. L'estimació dels models multiequacionals

Aquests models s'estimen per tres mètodes:

1) Els **mètodes directes** valoren cada equació per separat, sense tenir en compte que l'equació forma part d'un model multiequacional. El mètode més conegut és el d'MCO.

2) Els **mètodes d'informació limitada** valoren cada equació per separat, però tenen en compte informació addicional a l'equació estimada (registrada en la resta de les equacions del model). És a dir, consideren si una variable explicativa és endògena o exògena, i també les variables que no estan incloses en l'equació però que sí estan presents en el model. Els mètodes més habituals són el de mínims quadrats indirectes (MCI), el de mínims quadrats en dues etapes (MC2E), el de variables instrumentals (VI) i el de màxima versemblança d'informació limitada (MVIL).

3) Finalment, els **mètodes d'informació completa**, que estimen de manera conjunta totes les equacions del model i per tant tenen en compte tota la informació del model.

Els més habituals són el de mínims quadrats en tres etapes (MC3E) i el de màxima versemblança d'informació completa (MVIC).

En el quadre següent apareixen les propietats asimptòtiques dels estimadors MCO dels diferents models multiequacionals, sempre que es compleixin les hipòtesis bàsiques de l'MRLM en cada equació:

Model multiequacional	Propietats asimptòtiques dels estimadors MCO
No relacionat	Consistència i eficiència
Aparentment no relacionat	Consistència i no-eficiència
Recursiu	Consistència i no-eficiència
Integrat	Consistència i no-eficiència

Així, els models multiequacionals no relacionats no tindran problemes. Vegem què passa amb la resta dels models:

1) En els models aparentment no relacionats, els estimadors seran ineficients perquè l'estimació uniequacional no tindrà en compte tota la informació associada al model i haurem d'estimar per mètodes d'informació completa, com el **mètode d'estimació de Zellner**, que bàsicament consisteix a estimar el model multiequacional complet per MCG.

2) En els models integrats, els estimadors seran esbiaixats i inconsistents a causa de la presència d'altres variables endògenes correlacionades amb el terme de pertorbació com a variables explicatives. A més a més, seran ineficients perquè no consideren tota la informació disponible en el model. En aquest segon model, els estimadors d'informació limitada seran consistents, però ineficients, i els d'informació completa, consistents i eficients.

Els tres mètodes d'informació limitada més utilitzats són:

Aquest mètode és molt útil en equacions exactament identificades.

1) El mètode dels mínims quadrats indirectes (MCI) consisteix a estimar per MCO els paràmetres Π de la forma reduïda del model i, posteriorment, obtenir les estimacions dels paràmetres de la forma estructural ($Bi\Gamma$), a partir de la següent relació: $\Pi = -B^{-1}\Gamma \Rightarrow B\Pi + \Gamma = 0$.

Aquest mètode és útil tant en equacions sobreidentificades com en les exactament identificades.

2) El mètode dels mínims quadrats en dues etapes (MC2E) consisteix en primer lloc a estimar la forma reduïda per MCO per a obtenir una estimació de les variables endògenes ($\hat{Y}_t$): $\hat{Y}_t = \hat{\Pi} Z_t$. Després es torna a valorar l'equació estructural inicial per MCO, després d'haver substituït les variables endògenes presents com a variables explicatives (Y_t) pels seus valors estimats ($\hat{Y}_t$).

3) Per acabar, el mètode de les variables instrumentals (VI) consisteix a definir una matriu W d'instruments i aplicar la fórmula presentada en l'estimació dels models di-

nàmics. Com que s'utilitzen les mateixes variables com a instruments de les variables predeterminades d'una equació, podem trobar-nos amb el següent:

a) En el cas d'una equació exactament identificada, si com a instruments de les variables endògenes explicatives utilitzem les variables predeterminades omeses en l'equació, però presents en altres equacions del model, els resultats equivaldran als dels MCI.

b) D'altra banda, en el cas d'equacions sobreidentificades, si com a instruments utilitzem les obtingudes en la primera etapa dels MC2E, els estimadors VI equivaldran als MC2E.

1.2.5. Interpretació dels paràmetres del model

Una vegada estimat el model, hem d'interpretar el significat dels paràmetres. En aquest sentit, els paràmetres de la forma estructural únicament registren els efectes directes entre les variables explicatives i la variable endògena. Amb la finalitat de registrar els efectes directes i indirectes contemporanis (variacions en la variable endògena davant variacions en la variable exògena referides totes al mateix moment del temps), s'han de calcular els paràmetres de la forma reduïda. Pea acabar, per registrar tots els efectes contemporanis, així com aquells que fan referència a qualsevol moment del temps, hem d'analitzar els paràmetres de la forma final.

2. Model lineal generalitzat

2.1. Motivació

Fins ara, hem assumit que la relació entre la variable dependent i els regressors seguia la següent formulació matricial:

$$Y = X\beta + e$$

$$E(Y) = X\beta$$

Aquesta formulació pot ser limitada en el cas que la distribució dels errors no segueixi una distribució normal, i això succeeix, especialment, quan la variable dependent és de naturalesa qualitativa (és a dir, si es refereixen a atributs o categories, i no a valors numèrics), o bé si és una variable de recompte, que són les variables les observacions de les quals són nombres enters positius.

Analíticament, el **model lineal generalitzat** (GLM en la sigla en anglès) és una generalització flexible del model de regressió lineal ordinari, que permet la inclusió de variables dependents que generen distribucions de l'error diferents d'una distribució normal. El GLM generalitza la regressió lineal en permetre que el model lineal estigui relacionat amb la variable dependent a través d'una *funció d'enllaç*, i en permetre que la magnitud de la variància de cada mesurament sigui una funció del seu valor predit. Això és, la nova formulació té la següent forma:

Acrònims en diversos idiomes

El model lineal generalitzat rep l'acrònim **MLG** en català, però és més comú veure'l escrit com a **GLM** (*generalized linear model*), que és la sigla corresponent en anglès.

$$E(Y) = \mu = g^{-1}(X\beta)$$

On tenim els següents components:

1) Funció de distribució. Es tracta d'una funció de distribució que pertany a la família exponencial, a la qual pertanyen moltes de les distribucions més comunes, com la normal, exponencial, gamma, khi quadrat, beta, Dirichlet, Bernoulli, categòrica i Poisson. Això permet adaptar el model de regressió a la distribució específica de la variable dependent i, lògicament, l'error del model.

2) Mitjana de la funció de distribució (μ). Coincideix amb el valor esperat de Y, de manera que $E(Y) = \mu$.

3) Predictor lineal (η). És la magnitud que incorpora la informació sobre les variables explicatives en el model. Així, η s'expressa com una combinació lineal dels paràmetres desconeguts β, de manera que $\eta = X\beta$.

4) Funció d'enllaç (g). Aquesta funció defineix la relació entre el predictor lineal (η) i la mitjana de la funció de distribució (μ), de manera que $\eta = g(\mu)$.

5) Funció de la mitjana. Es basa a invertir la funció d'enllaç g, de manera que la mitjana sigui el predictor lineal de la funció d'enllaç, això és, $\mu = g^{-1}(\eta)$.

Així, el model de regressió lineal que hem vist fins ara es pot considerar com un cas particular de GLM, en què la funció de distribució és la normal. Això és:

$$E(Y) = \mu = X\beta$$

$$\eta = X\beta = \mu$$

Per a valorar un GLM no serveix la tècnica de mínims quadrats ordinaris, sinó que calen tècniques d'estimació més avançades i complexes, com el mètode de màxima versemblança.

2.2. Models lògit, pròbit i Poisson

Dins de la multitud de combinacions que ofereix la família exponencial, en aquesta secció repassarem els dos models més utilitzats per al cas de variables dependents qualitatives (lògit i pròbit), i el model bàsic per a analitzar dades de recompte (Poisson).

Les variables qualitatives es caracteritzen perquè els seus valors no són magnituds, sinó categories. En el cas més simple, les variables qualitatives dicotòmiques solament prenen dos valors (sí o no, negre o blanc, home o dona, etc.). Per a estudiar aquestes variables i explicar-ne el comportament en funció d'altres variables (això és, estimar un model de regressió) és molt útil codificar-les en forma de zeros i uns. No obstant això, és fonamental tenir molt clar que aquests valors 0 i 1 no són magnituds, sinó que representen dues categories diferents.

D'una banda, el **model de regressió lògit** es basa en la *funció lògit*, que és la inversa de la funció logística. D'aquesta manera, el *lògit* d'un nombre p entre 0 i 1 es defineix mitjançant la fórmula següent:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p).$$

Raó de probabilitat

Si p és una probabilitat, llavors $p/(1-p)$ és la raó de probabilitat corresponent (*odds* en anglès), i el lògit de la probabilitat és el logaritme de les probabilitats.

Auesta funció serveix de base per al *model de regressió lògit*, que té l'expressió siguiente:

$$E(Y) = \mu = g^{-1}(X\beta)$$

$$\eta = X\beta = \ln\left(\frac{\mu}{1-\mu}\right)$$

$$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)} = \frac{1}{1+\exp(-X\beta)}$$

Model de regressió lògit

El model lògit va ser introduït per Joseph Berkson el 1944.

És fonamental tenir en compte que la interpretació dels coeficients estimats d'un model lògit o pròbit és probabilística, i difereix de la interpretació que es dóna a un model de regressió lineal. En l'exemple aplicat que segueix a continuació veurem com s'interpreten els coeficients.

El **model de regressió pròbit** és molt similar al model anterior, i la diferència és que la distribució de referència és la normal, en lloc de la logística. En estadística, la funció pròbit és la funció quantil, és a dir, la funció de distribució acumulativa inversa (CDF, en la sigla en anglès), associat amb la distribució normal estàndard. D'aquesta manera, en el model de regressió pròbit la funció d'enllaç g adquireix la forma següent:

Model de regressió pròbit

La idea de pròbit va ser publicada el 1934 per Chester Ittner Bliss en un article sobre la forma de tractar les dades sobre el percentatge d'una plaga eliminada per un pesticida.

$$g = \Phi^{-1}(p)$$

Φ és la funció de distribució acumulativa de la distribució normal:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

El nom *pròbit* sorgeix d'ajuntar les paraules *probability unit*.

Els dos models se solen estimar usant tècniques de màxima versemblança.

Finalment, el **model de regressió de Poisson** s'empra quan la variable dependent és de recompte, és a dir, és el recompte d'ocurrències d'un esdeveniment en un espai de temps determinat, i pren valors enters i positius. En aquest cas, la distribució de referència és la distribució de Poisson, i el model adquireix la forma següent:

La distribució de Poisson

Aquesta distribució deu el nom a Siméon Denis Poisson (1781-1840), matemàtic i físic francès, que la va donar a conèixer el 1838 en el seu treball *Recherches sur la probabilité des jugements en matières criminelles et matière civile*.

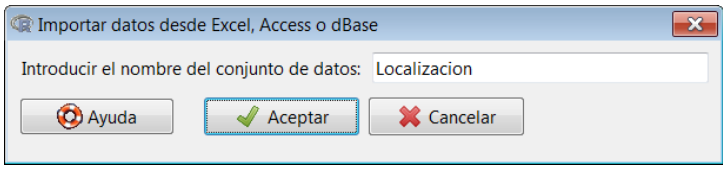
$$E(Y) = \mu = g^{-1}(X\beta)$$

$$\eta = X\beta = \ln(\mu)$$

$$\mu = \exp(X\beta) = \exp(\eta)$$

2.3. Aplicació empírica

L'objectiu d'aquesta secció és presentar una aplicació dels models fins ara vistos amb R-Commander. L'estudi fa referència a la localització de noves empreses en els municipis de Catalunya en el període 2001-2004. Com hem fet fins ara, el primer pas és importar les dades d'un full de càlcul i crear-ne un conjunt, al qual denominarem *localizacion*:



Visualitzant les dades, veiem les variables que en componen el conjunt:

Localizacion

	MUNICIPIO	LOC	LOC_DICO	DIM	TASA_ACT	EDU	T_CAP	T_AERO
1	08001-Abrera	13	1	14.570946	50.01160	8.4082	57	22
2	08002-Aguilar de Segarra	0	0	11.250000	40.90909	8.1936	74	58
3	08003-Alella	0	0	18.084906	45.26564	11.0880	64	25
4	08004-Alpens	0	0	18.000000	45.32374	9.1787	97	73
5	08005-Ametlla del Vallès (L')	4	1	13.148649	47.59498	10.9050	66	31
6	08006-Arenys de Mar	9	1	12.995370	43.73977	9.2117	70	38
7	08007-Arenys de Munt	22	1	15.458333	44.53113	8.8717	71	39
8	08008-Argençola	0	0	14.500000	45.31250	9.6831	69	53
9	08009-Argentona	9	1	13.993846	45.95796	9.2986	66	31
10	08010-Artés	9	1	13.328947	44.82301	7.9375	72	47

La descripció de les variables és la següent:

MUNICIPIO: nom del municipi i codi postal.

LOC: nombre d'empreses que s'han localitzat en el municipi.

LOC_DICO: variable dicotòmica que pren un valor d'1 si s'han localitzat noves empreses ($LOC > 0$) i el valor 0 si no se n'ha localitzat cap ($LOC = 0$).

DIM: dimensió mitjana de les empreses del municipi, expressada en nombre de treballadors.

TASA_ACT: taxa d'activitat del municipi, expressada com el quocient del nombre de treballadors i de la població total municipal.

EDU: mitjana d'anys d'educació de la població municipal.

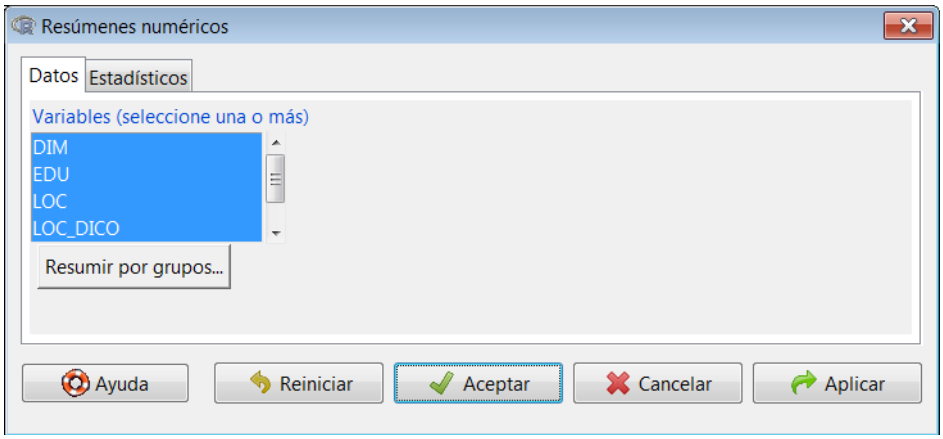
T_CAP: temps mitjà de transport a la capital més propera, en minuts.

T_AERO: temps mitjà de transport a l'aeroport, en minuts.

Abans de fer l'anàlisi empírica, sempre és útil calcular els estadístics bàsics de les variables, cosa que es fa accedint a la ruta següent del menú desplegable:

Estadístics / Resums / Resums numèrics

Seleccionem totes les variables en el quadre de diàleg resultant:



El resultat en la consola és el següent:

```
> numSummary(Localizacion[,c("DIM", "EDU", "LOC", "LOC_DICO",
+ "T_AERO", "T_CAP", "TASA_ACT")], statistics=c("mean", "sd",
+ "quantiles"), quantiles=c(0,.25,.5,.75,1))

            mean      sd     0%    25%    50%     75%     100%     n
DIM        15.20    8.54   0.00  10.57  13.28   17.79   120.00   941
EDU         8.49    1.01   4.23   7.82   8.42    9.12    11.99   941
LOC         3.55   12.42   0.00   0.00   0.00    2.00   201.00   941
LOC_DICO    0.44    0.49   0.00   0.00   0.00    1.00     1.00   941
T_AERO     49.11   33.00   0.00  27.00  41.00   63.00   190.00   941
T_CAP      87.41   23.30  56.00  70.00  82.00  101.00   190.00   941
TASA_ACT   43.76    4.76  23.46  40.67  44.13   46.99    58.24   941
```

Veiem que el valor mitjà de la variable DIM és d'uns 15 treballadors per empresa, i que s'han localitzat empreses en el 44% dels municipis (ja que la mitjana de *LOC_DICO* és $0,44$).

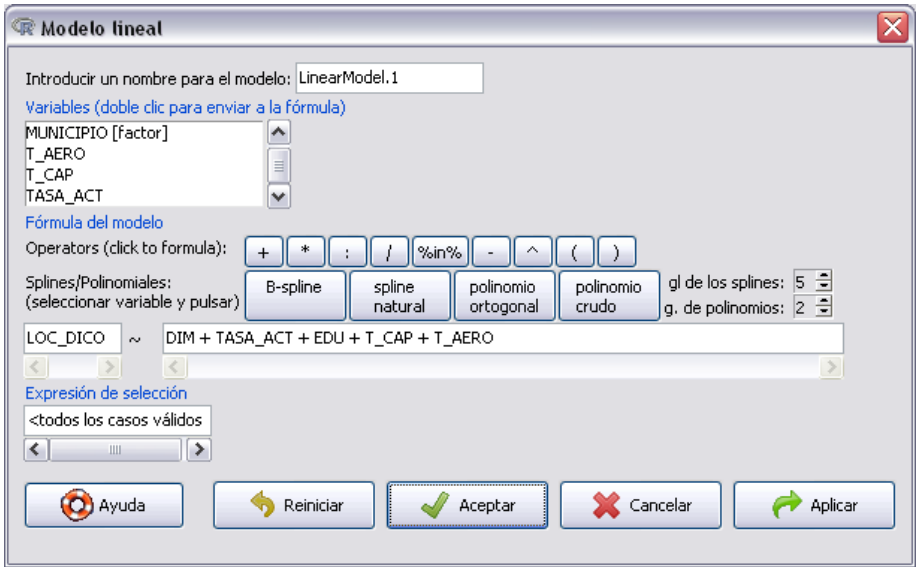
En el primer model que analitzarem, la variable dependent serà la variable dicotòmica *LOC_DICO*, això és, estudiarem quin efecte tenen les variables explicatives sobre la localització de noves empreses en un municipi. Analíticament, prenem la forma funcional següent:

$$LOC\_DICO = f(DIM,\ TASA\_ACT,\ EDU,\ T\_CAP,\ T\_AERO)$$

El primer pas serà fer una estimació per mínims quadrats ordinaris (OLS en la sigla en anglès). Per fer-la, acudim a la ruta següent del menú desplegable:

Estadístics / Ajust de models / Model lineal

En el quadre de diàleg resultant, introduïm la relació entre les variables, com segueix:



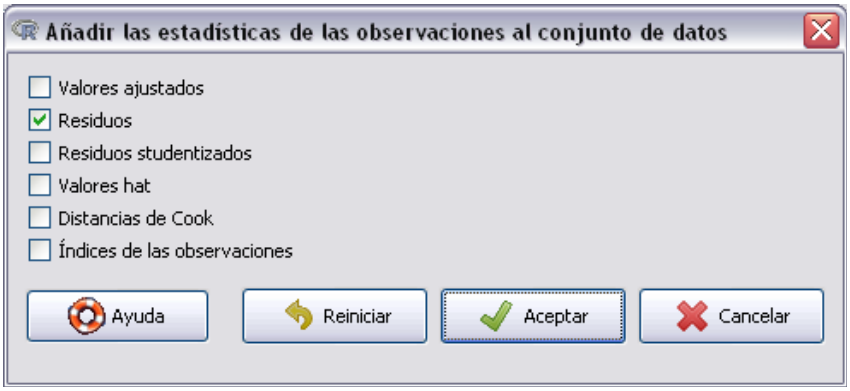
El resultat és el següent:

```
> LinearModel.1 <- lm(LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP +
+ T_AERO, data=Localizacion)

> summary(LinearModel.1)

Call:
lm(formula = LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    data = Localizacion)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04317 -0.38490 -0.05569  0.38643  1.10766

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept)  0.6640403  0.1743117   3.809 0.000148 ***
DIM         -0.0119723  0.0017098  -7.002 4.81e-12 ***
TASA_ACT     0.0236518  0.0033335   7.095 2.55e-12 ***
EDU         -0.0471142  0.0156403  -3.012 0.002662 **
T_CAP       -0.0084925  0.0009044  -9.391  < 2e-16 ***
T_AERO       0.0014869  0.0006411   2.319 0.020589 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4417 on 935 degrees of freedom
Multiple R-squared:  0.216, Adjusted R-squared:  0.2118
F-statistic: 51.52 on 5 and 935 DF,  p-value: < 2.2e-16
```

Com observem, encara que l'ajust del model sigui més aviat pobre, tots els coeficients de les variables explicatives són significatius. A ra bé, quina validesa té aquesta estimació, tenint en compte la naturalesa dicotòmica de la variable dependent? D'una banda, en una estimació per OLS les prediccions del model no estaran necessàriament entre zero i u. A més a més, els errors es distribueixen com una distribució bimodal, i no com una normal. Per veure això, primer extraurem els residus de l'estimació que acabem de fer, i els afegirem a les observacions al conjunt de dades. Anant a la ruta següent:

Models / Afegir les estadístiques de les observacions al conjunt de dades

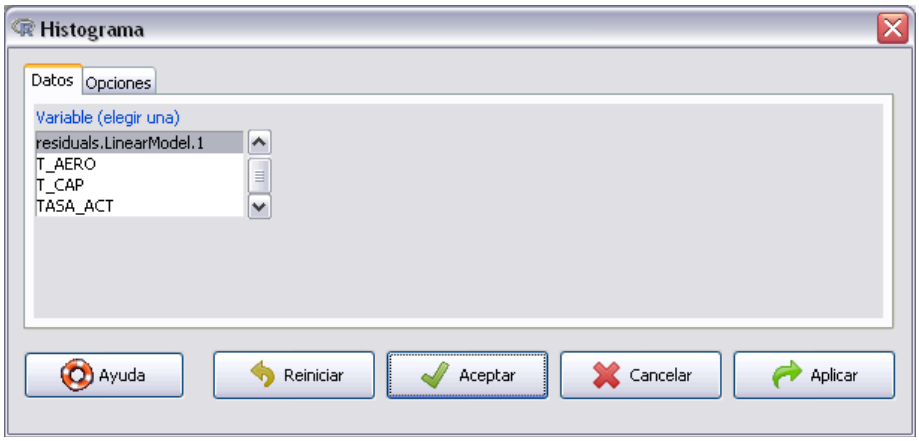
Ens apareix el quadre de diàleg on hem de seleccionar quines magnituds derivades de l'estimació volem emmagatzemar com a variables:



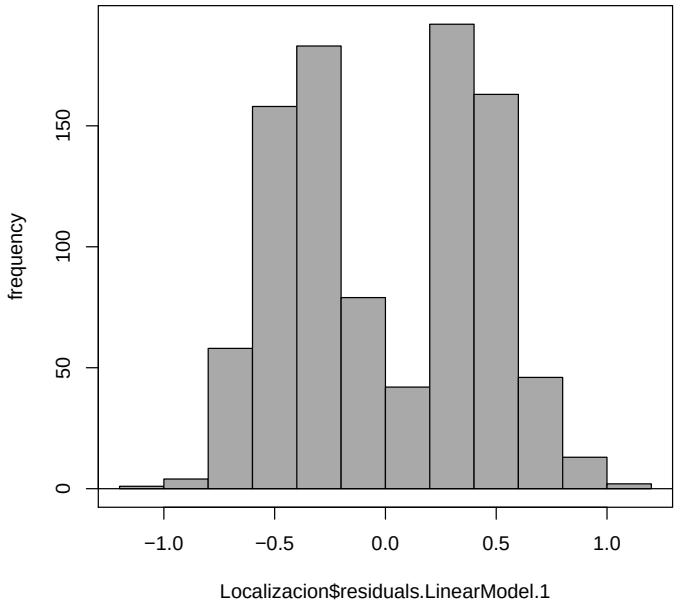
Una vegada seleccionats els residus, en veurem l'histograma accedint a la ruta següent:

Gràfiques / Histograma

Seleccionem els residus entre totes les variables, com segueix:



I obtenim l'histograma dels residus:



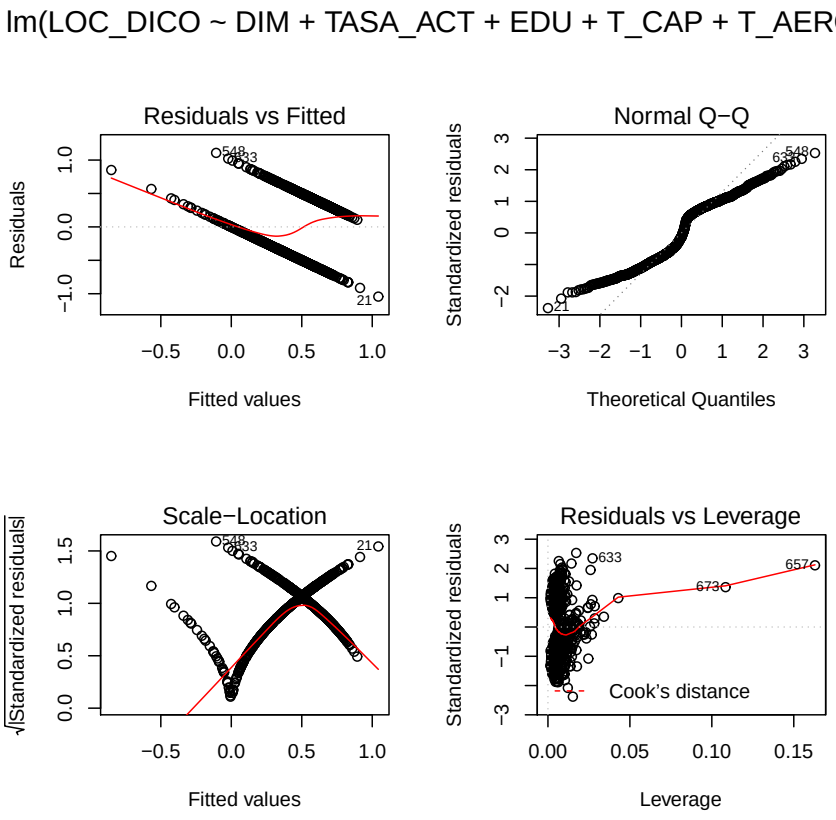
Distribució bimodal

En estadística, una distribució bimodal és una distribució de probabilitat contínua amb dues modes diferents, que apareixen com dos pics diferents (màxims locals) en la funció de densitat de probabilitat.

Com veiem, la distribució és bimodal, i no té la forma característica d'una distribució normal, cosa que té implicacions importants a l'hora de fer inferència. Podem, alternativament, comprovar el comportament particular dels residus si obtenim el gràfic compost del diagnòstic del model:

Models / Gràfics bàsics de diagnòstic

Aquí podem veure el resultat:



No obstant això, el principal problema de l'estimación MCO és que pressuposa que l'efecte marginal d'un augment d'una variable explicativa sobre la variable dependent és el coeficient, això és, per a una mostra de $i = 1, \ldots, n$ observacions i k variables explicatives, l'efecte marginal d'un augment del regressor x_j sobre la variable dependent és:

$$\frac{\partial y_i}{\partial x_{ji}} = \beta_j$$

Aquest és un supòsit molt restrictiu, ja que la variable dependent en un model binari està acotada entre 0 i 1, i s'interpreta probabilísticament. Dit d'una altra manera, l'increment de l'efecte d'un augment de x_j sobre l'increment de y no pot ser el mateix per a valors alts i baixos de x_j. A més a més, la relació real no és lineal, ja que s'espera un efecte molt baix d'un canvi de x_j sobre y quan x_j és, o bé molt baix, o molt alt.

El *model de regressió lògit* soluciona aquest problema, ja que el model treballa explícitament en termes de probabilitats. Això és, per a una mostra de $i = 1, \ldots, n$ observacions i k variables explicatives, tenim el següent:

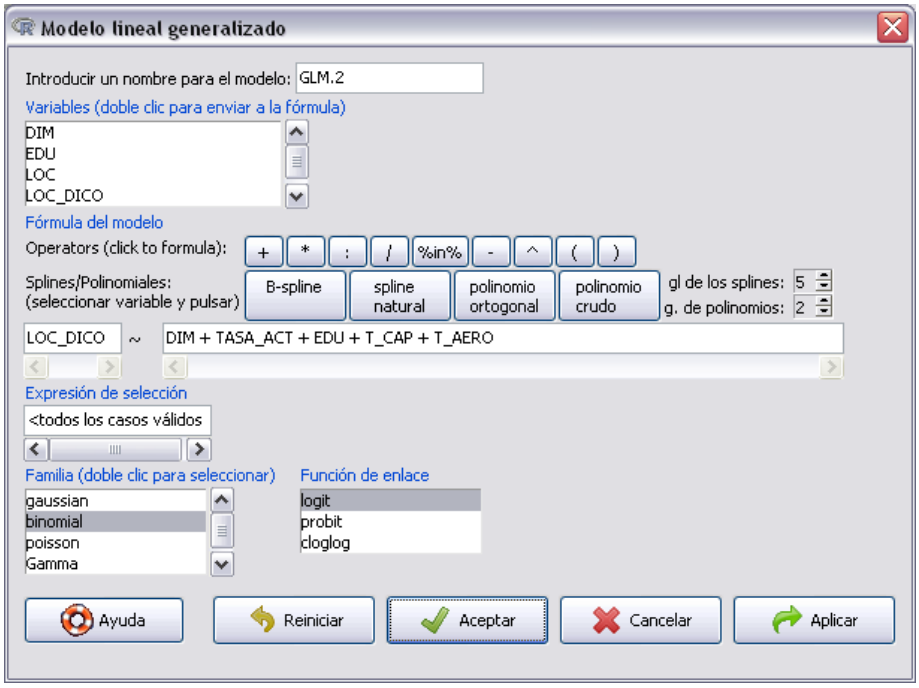
$$\text{logit}(P_i) = log\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

On P_i és la probabilitat d'ocurrència de l'esdeveniment (en el nostre cas, la localització d'empreses) i $\left(\frac{P_i}{1-P_i}\right)$ és la raó de probabilitat o *odds* en la terminologia anglesa, i s'interpreta de la manera següent: si la probabilitat que hi hagi localitzacions empresarials és de $p = 0,75$, la probabilitat que no n'hi hagi és d'$1 - p = 0,25$, de manera que la raó de probabilitat serà $0,75/0,25 = 3$, això és, la probabilitat que hi hagi localitzacions és de 3 a 1 a favor que sí que n'hi hagi.

Per estimar aquest model amb R-Commander, anem a la ruta següent del menú desplegable:

Estadístics / Ajust de models / Model lineal generalitzat

En el quadre de diàleg resultant, a més d'introduir la variable dependent i las variables independent, introduirem la família (*binomial*) i la funció d'enllaç (*logit*):



El resultat que obtenim és el següent:

```
> GLM.2 <- glm(LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO
    , family=binomial(logit), data=Localizacion)

> summary(GLM.2)

Call:
glm(formula = LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    family = binomial(logit), data = Localizacion)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5610  -0.9145  -0.3556   0.9402   2.8054

Coefficients:
             Estimate Std. Error z value Pr(>z)
(Intercept)  1.866726   0.970452   1.924  0.05441 .
DIM         -0.095450   0.013737  -6.949 3.69e-12 ***
TASA_ACT     0.120666   0.018722   6.445 1.16e-10 ***
EDU         -0.275304   0.083802  -3.285  0.00102 **
T_CAP       -0.045395   0.005025  -9.034  < 2e-16 ***
T_AERO       0.004852   0.003555   1.365  0.17232
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1294.1  on 940  degrees of freedom
Residual deviance: 1044.3  on 935  degrees of freedom
AIC: 1056.3

Number of Fisher Scoring iterations: 5
```

Com hem d'interpretar el resultat d'aquesta estimació? A diferència de l'MRLM, en un model binari com el lògit la influència que tenen les explicatives sobre la probabilitat de triar l'opció donada per $y_i = 1$ no depèn simplement del valor dels coeficients, sinó també del valor que prenen les variables explicatives, és a dir, $\frac{\partial y_i}{\partial x_{ji}} = f(\beta_j, x_{ji}$. Encara que la interpretació sigui complexa, podem partir de la base que, en incrementar una unitat la variable x_j, la probabilitat P_i passa a ser $P_i^{'}$, de manera que:

$$\text{logit}(P_i^{'}) = \beta_j + \text{logit}(P_i)$$

Si reescrivim aquesta expressió, arribem a una fórmula que ens serà molt útil a l'hora d'interpretar el resultat:

$$\frac{\frac{P_i^{'}}{1-P_i^{'}}}{\frac{P_i}{1-P_i}} = \exp(\beta_j)$$

Això és, $\exp(\beta_j)$ es denomina l'**odds ratio**, i indica el canvi relatiu que experimenta el quocient de probabilitats quan la variable x_j augmenta una unitat. En el nostre exemple, el coeficient estimat de la variable *temps de transport a les capitals (T_CAP)* és de $\hat{\beta}_4 = -0,045$, això és, $e^{-0,045} = 0,955$, cosa que indica que l'augment d'1 km en la distància a les capitals provoca una petita disminució del quocient de probabilitats. La variable *anys d'educació (EDU)* té un coeficient estimat negatiu de magnitud més gran ($\hat{\beta}_3 = -0,27$), de manera que un augment marginal d'aquesta variable (un any més d'escolarització) causa un descens en el quocient de probabilitats de $e^{-0,27} = 0,76$, és a dir, es redueix aproximadament en una quarta part.

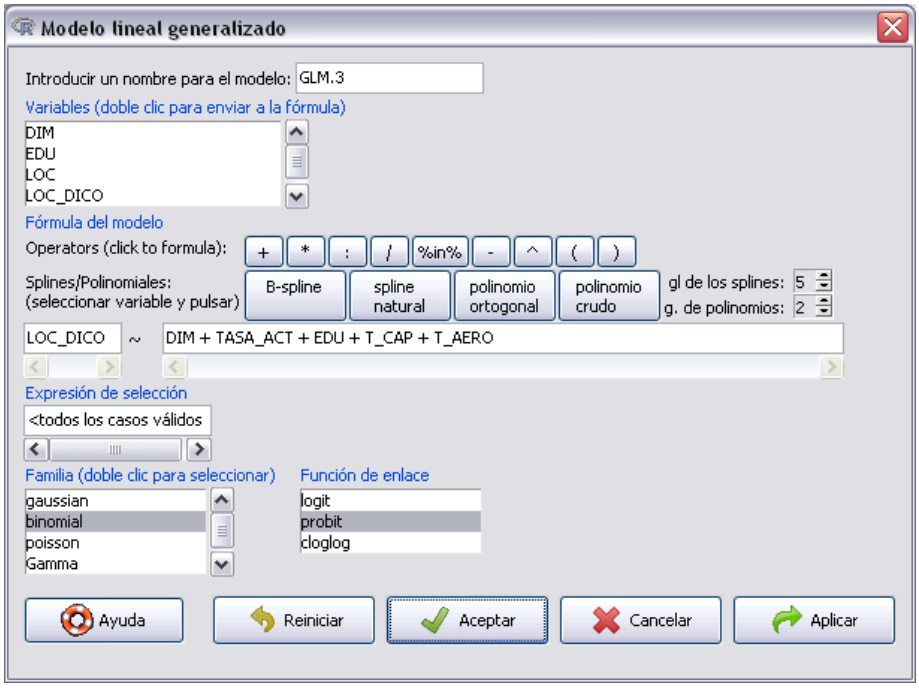
Odds ratio

No hi ha unanimitat sobre la traducció d'aquest concepte al català, per això aquí el deixem en l'expressió anglesa. Algunes propostes de traducció són *raó de momios, raó relativa, raó d'oportunitats, raó de productes encreuats, raó de desigualtats, raó de disparitat, raó d'excés* o, simplement, *raó d'odds*.

El *model de regressió pròbit*, como s'ha comentat abans, és molt similar al model anterior, i es diferencia bàsicament en el fet que el model pròbit es basa en una distribució normal acumulada. Per a estimar aquest model amb R-Commander, es procedeix d'una manera molt similar al cas anterior:

Estadístics / Ajust de models / Model lineal generalitzat

En el quadre de diàleg resultant, a més d'introduir la variable dependent i les variables independents, introduirem la família (*binomial*) i la funció d'enllaç (*probit*):



El resultat és el següent:

```
> GLM.3 <- glm(LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO
    ,
+ family=binomial(probit), data=Localizacion)

> summary(GLM.3)

Call:
glm(formula = LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    family = binomial(probit), data = Localizacion)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6358  -0.9327  -0.3466   0.9499   2.9670

Coefficients:
             Estimate Std. Error z value Pr(>z)
(Intercept)  1.030755   0.571132   1.805  0.07111 .
DIM         -0.053994   0.007728  -6.987 2.81e-12 ***
TASA_ACT     0.070029   0.010875   6.440 1.20e-10 ***
EDU         -0.154390   0.049434  -3.123  0.00179 **
T_CAP       -0.026678   0.002923  -9.128  < 2e-16 ***
T_AERO       0.003247   0.002085   1.557  0.11944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1294.1  on 940  degrees of freedom
Residual deviance: 1048.4  on 935  degrees of freedom
AIC: 1060.4

Number of Fisher Scoring iterations: 5
```

Com veiem, el resultat de l'estimació, tant els coeficients individuals com l'ajust, és molt similar a l'estimació anterior. L'estadístic AIC (*akaike information criterion*), molt utilitzat en l'estimació per màxima versemblança, és lleugerament menor per a l'estimació *logit*, de manera que ens quedaríem amb l'estimació del model anterior.

Akaike information criterion (AIC)

És una mesura de la *qualitat relativa* d'un model economètric respecte a d'altres, i per a un conjunt específic de dades. És a dir, és una mesura que serveix per a comparar diversos models, un *trade-off* entre la bondat d'ajust del model i la complexitat d'aquest (mesura com la quantitat de paràmetres que s'han de valorar).

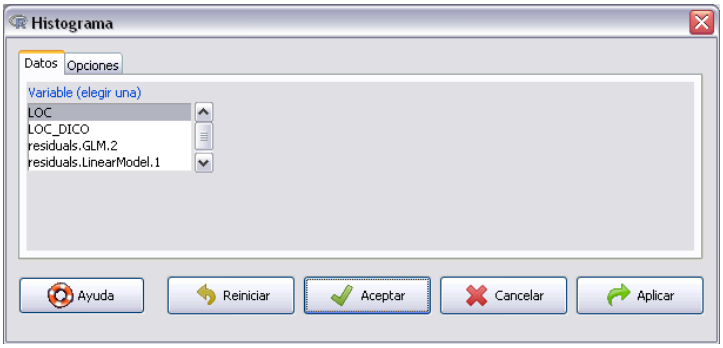
Per acabar, veurem un exemple de l'estimació d'un *model de regressió de Poisson*. Per a això, estimarem un model de regressió com l'anterior, però ara prenent com a variable dependent el *nombre d'empreses localitzades (LOC)*:

$$LOC = f(DIM,\ TASA\_ACT,\ EDU,\ T\_CAP,\ T\_AERO)$$

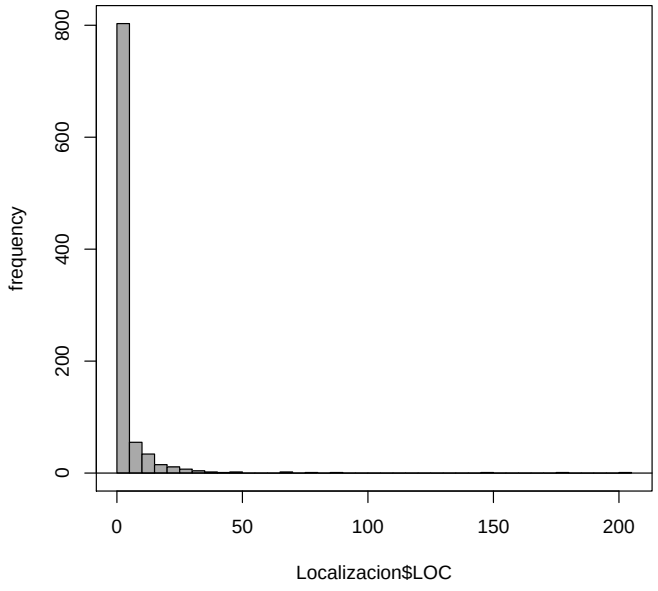
D'aquesta manera, ara la variable dependent és quantitativa, però amb una distribució peculiar. Vegem-ne l'histograma anant a l'opció següent del menú desplegable:

Gràfiques / Histograma

Seleccionem la variable *LOC* en el quadre de diàleg:



Obtenim el gràfic següent:



Com veiem, la distribució de la variable dista molt de ser normal. La majoria de les observacions estan situades en els primers valors $(0, 1, 2, \ldots)$, i el nombre de recomptes més grans de 40 és molt, molt petit. De fet, la mitjana de la variable LOC és de $3,55$. Aquesta distribució, doncs, es pot associar més aviat a una Poisson. Aquest model té la formulació següent:

$$E(Y) = \mu = g^{-1}(X\beta)$$

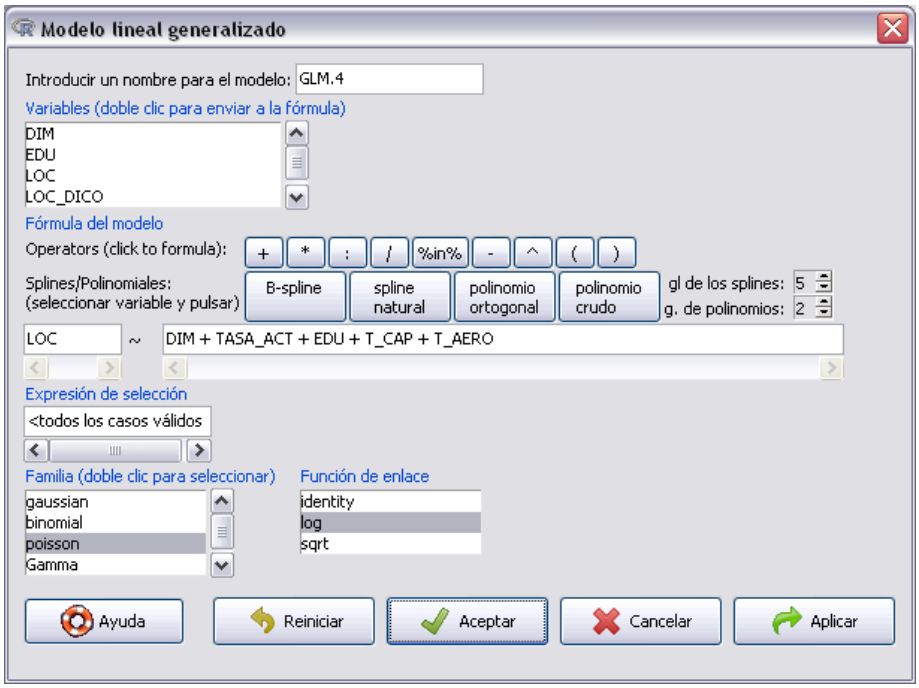
$$\eta = X\beta = \ln(\mu)$$

$$\mu = \exp(X\beta)$$

Per estimar el model, anem a la ruta següent del menú desplegable:

Estadístics / Ajust de models / Model lineal generalitzat

En el quadre de diàleg, introduïm les variables del model i la família Poisson:



El resultat és com segueix:

```
> GLM.4 <- glm(LOC ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    family=poisson(log), data=Localizacion)

> summary(GLM.4)

Call:
glm(formula = LOC ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    family = poisson(log),
```

```
    data = Localizacion)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.1733  -1.8274  -0.8423   0.0227  28.0688

Coefficients:
             Estimate Std. Error z value Pr(>z)
(Intercept)  6.549514   0.268266  24.414   <2e-16 ***
DIM         -0.102111   0.003981 -25.649   <2e-16 ***
TASA_ACT     0.061121   0.004840  12.629   <2e-16 ***
EDU         -0.175997   0.020962  -8.396   <2e-16 ***
T_CAP       -0.062467   0.001665 -37.526   <2e-16 ***
T_AERO      -0.011990   0.001207  -9.931   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11596.2  on 940  degrees of freedom
Residual deviance:  6948.9  on 935  degrees of freedom
AIC: 8292.3

Number of Fisher Scoring iterations: 7
```

La interpretació dels paràmetres estimats depèn de quin tipus de variable explicativa tingui associada. Així, per al coeficient estimat $\hat{\beta}_k$, podem tenir tres casos:

1) *Variable dummy*: la mitjana condicional de la variable dependent serà $\exp(\hat{\beta}_k)$ vegades més gran si la variable independent x_k pren el valor 1 en lloc de 0.

2) *Variable contínua*: el coeficient associat $\hat{\beta}_k$ és una semielasticitat, de manera que $100 \cdot \hat{\beta}_k$ és el canvi percentual en la mitjana de la variable dependent quan la variable explicativa augmenta en una unitat.

3) *Variable en logaritmes*: el coeficient associat $\hat{\beta}_k$ és una elasticitat, i el canvi percentual en la mitjana de la variable dependent es produeix quan la variable explicativa augmenta en un 1%.

Variables dummy

Es tracta de variables explicatives qualitatives dicotòmiques, que adquireixen el valor 0 o 1 per indicar l'absència o la presència d'algun efecte categòric. També es coneixen com a variables indicador, variables de disseny o indicadors booleans.

3. Models amb dades de panel

3.1. Introducció

L'objectiu d'aquest capítol és proporcionar una introducció completa a les tècniques de dades de panel. El terme *panel de dades* fa referència a un conjunt de dades amb observacions temporals per als mateixos individus, cosa que permet a l'investigador seguir un mateix individu durant el temps. A causa de l'àmplia disponibilitat d'aquest tipus de dades, aquestes s'utilitzen en diferents camps, com l'economia laboral, l'anàlisi de la productivitat, la demografia i les finances, per exemple.

Sobre la terminologia

Mentre que en estadística i econometria es denomina **dades de panel** un estudi amb dades d'aquests individus a través del temps, en bioestadística aquest conjunt de dades i tècniques reben el nom de **dades i estudis longitudinals**.

Les principals característiques de l'anàlisi de dades de panel són les següents:

1) *Control de l'heterogeneïtat individual*: en economia és fonamental l'estudi d'agents econòmics heterogenis (països, persones, empreses, etc.). Els estudis solament amb dades *cross-section* o amb sèries temporals poden donar resultats esbiaixats.

2) *Augment de graus de llibertat:* les dades de panel impliquen mostres més grans, cosa que resulta en més eficiència en l'estimació.

3) *Reducció de multicol·linealitat*: la unió de les dimensions individual i temporal implica la inclusió d'una gran quantitat de variabilitat i de la informació. De fet, la variació de les dades es pot descompondre en la variació entre diferents individus (*between*) i la variació dins de cada individu (*within*).

4) *Estudi de fenòmens dinàmics:* els estudis de tall transversal produeixen una imatge estàtica del tema analitzat, mentre que les dades de panel poden descobrir diferents efectes temporals i analitzar l'evolució dinàmica de la imatge.

5) *Identificació d'efectes determinats*: hi ha certs efectes que solament es poden analitzar mitjançant estudis amb dades de panel.

6) *Més possibilitats de modelització:* per exemple, en els models de retards distribuïts, en haver-hi més variabilitat s'han d'imposar menys restriccions al model, cosa que millora l'anàlisi.

7) *Eliminació del biaix d'agregació:* les dades que agreguen individus, en general, presenten un biaix important.

No obstant això, obtenir dades de panel a vegades comporta dificultats i limitacions. Un problema és el disseny i la recollida de dades: falta de resposta, *missing values*, errors de mesurament, etc. Un altre problema és el del biaix de selecció: biaix mostral, autoselecció, pèrdua d'individus, dades censurades i truncades, etc. A més a més,

cal destacar que les dimensions del panel són rellevants per a les propietats asimptòtiques dels estimadors. Essent N el nombre d'individus i T el nombre d'observacions mostrals, es poden distingir dos casos generals:

1) $N > T$: es correspon amb dades microeconòmiques o àmplies mostres d'individus.

2) $T > N$: sol ser el cas de dades agregades, com regions o països, observats durant un llarg període de temps.

Analíticament, una mostra de dades de panel es pot descriure de la manera següent:

$$\{(y_{it}, x_{it}) : \ i = 1, \ldots, N; \ \ t = 1, \ldots, T\}.$$

Les observacions estan aparellades (y_{it}, x_{it}), el subíndex i fa referència als individus $(i = 1, \ldots, N)$ i t fa referència al període temporal $(t = 1, \ldots, T)$. Així doncs, la dimensió total de la mostra és $N \times T$. y_{it} és la variable dependent o explicada, mentre que x_{it} és el vector de variables explicatives o regressors. Aquest inclou k variables: $x_{1,it}, \ldots, x_{k,it}$.

Com amb qualsevol model economètric, hi ha dos passos fonamentals en l'anàlisi d'un model amb dades de panel:

1) *Especificació del model*: el fet d'assumir si s'inclouen efectes individuals o temporals, o tots dos, determinarà el resultat de l'estimació. El conjunt de regressors triat, a més de la forma funcional en què s'inclouen (lineal, additiva, etc.), també marcarà el resultat.

2) *Estimació del model:* una vegada especificat el model, hi ha diversos mètodes d'estimació disponibles, cada un amb una sèrie de propietats i característiques. El coneixement previ del fenomen analitzat i la disponibilitat de contrastos i anàlisis de la variància ajudaran a escollir correctament.

Depenent de les hipòtesis i assumpcions que fem del model que hem d'estudiar, hi ha tres possibles especificacions de partida (u_{it} és en tots els casos el terme de pertorbació):

1) *Model amb efectes individuals*:

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it}.$$

2) *Model amb efectes temporals*:

$$y_{it} = x_{it}'\beta + \eta_t + u_{it}.$$

3) *Model amb efectes individuals i temporals*:

$$y_{it} = x'_{it}\beta + \alpha_i + \eta_t + u_{it}.$$

Un estudi descriptiu previ de les variables ajudarà a triar una d'aquestes tres especificacions.

3.2. Estimació d'un model de dades de panel

3.2.1. Mínims quadrats ordinaris (MCO)

En nomenclatura anglesa, aquest estimador també es coneix com a *pooled OLS*. La hipòtesi bàsica d'aquest mètode d'estimació és suposar que els efectes individuals són comuns entre els individus ($\alpha_i = \alpha$). D'aquesta manera, el model que s'ha d'estimar té la forma:

Acrònims en diversos idiomes

L'estimació per mínims quadrats ordinaris rep l'acrònim **MCO** en català, però és més comú veure'l escrit com a **OLS** (*ordinary least squares*)„ que són la sigla corresponent en anglès. En aquest manual usarem els dos acrònims indistintament.

$$y_{it} = +\alpha + x'_{it}\beta + u_{it}.$$

Aquest mètode agrega les dues dimensions i i t, sense tenir en compte les possibles particularitats de cada individu. Les propietats d'aquest estimador dependran de la possible existència d'efectes individuals α_i correlacionats amb els regressors x_i:

1) $E(x_{it}u_i) \neq 0$: en aquest cas, OLS serà **esbiaixat** i **inconsistent**. La causa és que, com que α_i no és modelitzat, passa a formar part del terme de pertorbació u_i, amb la qual cosa **no** es compleix la condició d'ortogonalitat $E(x_{it}u_{it}) = 0$. Això passarà si, per exemple, α_i és una variable omesa.

2) $E(x_{it}u_i) = 0$: l'estimador OLS serà **consistent**, però **no serà eficient** a no ser que u_i sigui esfèric, és a dir, homoscedàstic i no correlacionat.

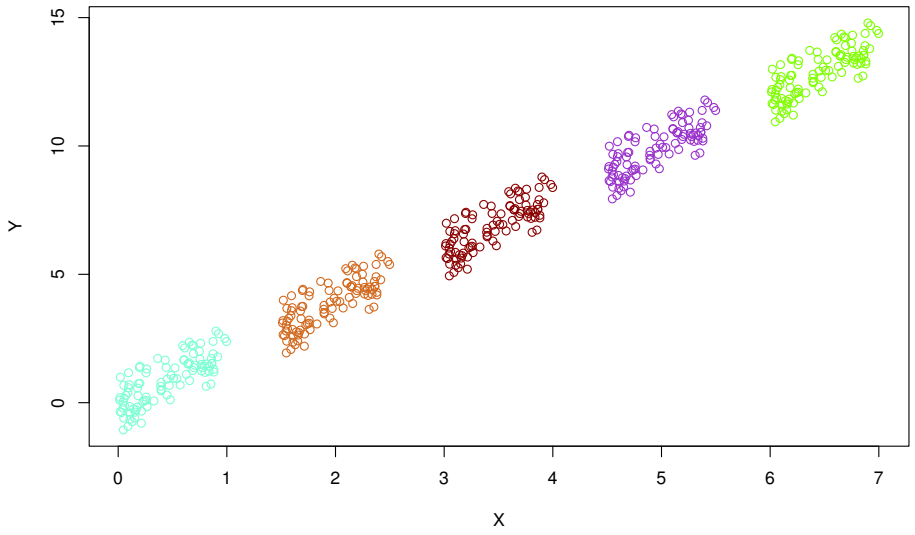
3) En general, u_i no és esfèric, l'estimació es pot millorar mitjançant mínims quadrats generalitzats (GLS).

Vegem un exemple fictici. Suposem el model següent:

$$C_{it} = \alpha + \beta R_{it} + u_{it}$$

On C és el consum i R és la renda disponible. La mostra està disponible per a $N = 5$ individus i $T = 100$ períodes temporals.

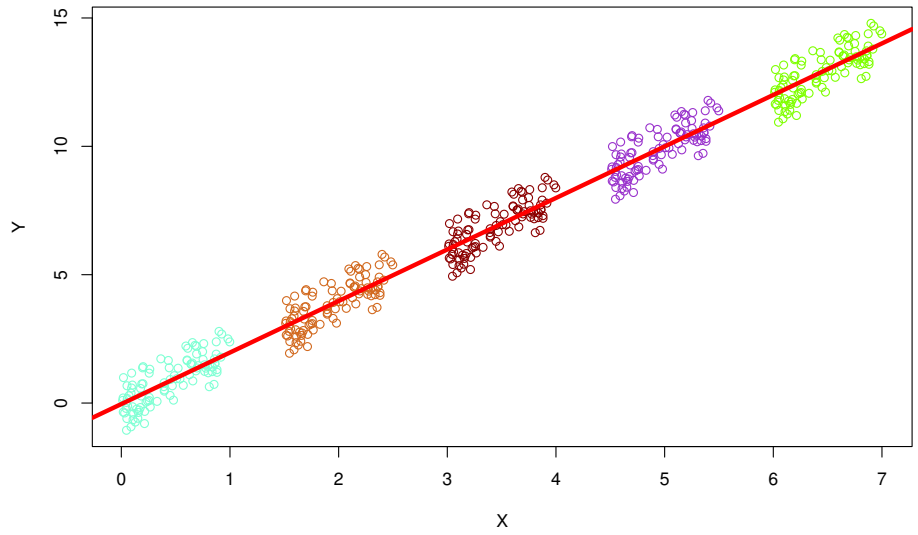
Una primera visualització del núvol de punts permet veure com els cinc individus (cada un mostrat amb un color diferent) mostren diferents nivells de renda i consum.



Una estimació mitjançant OLS permet obtenir el resultat següent:

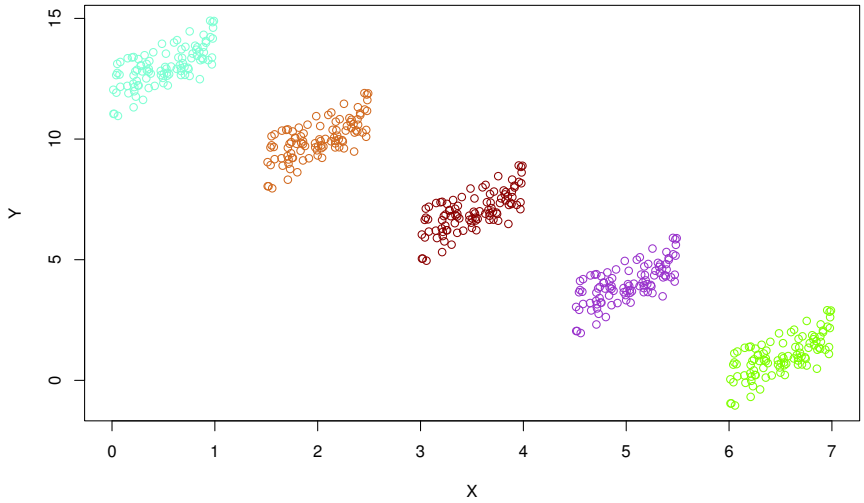
Variable	Coef.	Error est.	Estad. t	Valor-p
Constant $\hat{\alpha}$	$-0,004$	$0,050$	$0,080$	$0,93$
Renda ($\hat{\beta}$)	1.996	$0,012$	$165,260$	$0,00$
Estadístic F	27312			$0,00$
R^2	$0,97$			

Com es pot comprovar, l'ajust del model és molt bo. A més a més, el coeficient $\hat{\beta}$ és significatiu i positiu. El resultat, gràficament, és el següent:



Suposem ara que en l'exemple anterior hi ha una variable omesa: el patrimoni de cada individu. Suposem, a més a més, que els individus amb un alt patrimoni generen poca

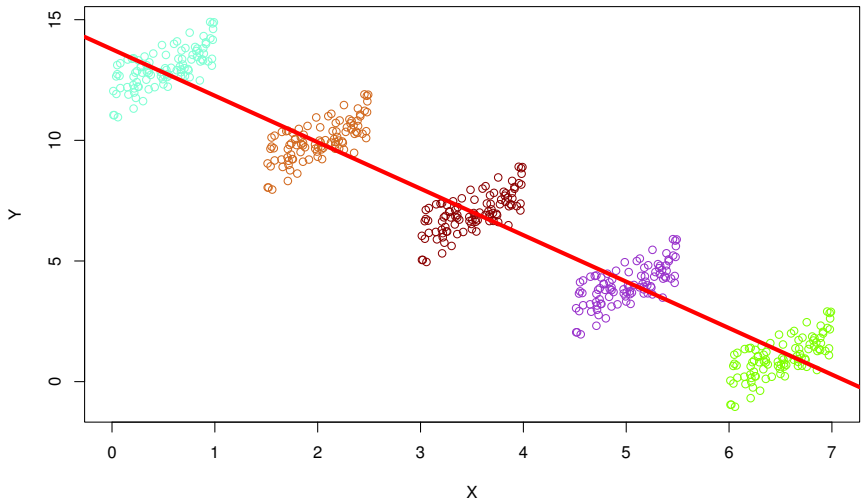
renda, però igualment mantenen un alt nivell de consum. El diagrama de dispersió adquireix la forma següent:



El resultat de l'estimador OLS en aquest nou escenari és el següent:

Variable	Coef.	Error est.	Estad. t	Valor-p
Constante $\hat{\alpha}$	$13,77$	$0,109$	$125,6$	$0,00$
Renta ($\hat{\beta}$)	$-1,92$	$0,026$	$-72,2$	$0,00$
Estadístico F	5223			$0,00$
R^2	$0,91$			

L'ajust del model continua essent molt bo, i els coeficients són significatius. On és el problema? Per començar, hem de preguntar-nos: és lògic que la renda exerceixi un efecte negatiu sobre el consum? Com veiem en el gràfic següent, el problema és que l'estimador OLS passa pel núvol de punts sense tenir en compte que, per a cada individu, el pendent individual és positiu.



3.2.2. Efectes fixos (WG/LSDV)

En anglès es denomina *within gropus* (WG) o *least squares dummy variables estimator* (LSDV). Aquest mètode és adequat quan els individus es representen a ells mateixos, i és possible que difereixin en característiques i comportament (regions, països, etc.). D'aquesta manera, diferències entre individus es reflectiran en diferències en la constant. Analíticament, s'assumeix que els efectes individuals són un conjunt de N paràmetres que s'han d'estimar.

Formalment, l'especificació d'efectes fixos adquireix la forma següent:

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}$$

$$\alpha_i = \alpha + \mu_i$$

La constant té dos elements: α és la constant comuna i μ_i és un efecte individual que ha de ser estimat. És important destacar que μ_i és un paràmetre que s'ha d'estimar i no és part del terme de pertorbació u_i. Això provoca que es compleixi la condició d'ortogonalitat $E(x_{it}u_{it}) = 0$, ja que μ_i no està recollit en u_i.

El mètode d'estimació té diverses alternatives.

1) Least squares dummy variables (LSDV): aquesta transformació implica incloure una variable *dummy* per a cada individu i. Si reescrivim el model en notació matricial, tenim:

$$Y_i = \iota_T \alpha_i + X_i\beta + u_i$$

On ι_T és un vector d'uns de dimensió $T \times 1$. Si s'agreguen tots els individus, s'obté:

$$Y = D\alpha + X\beta + u$$

En què D és una matriu de *dummies* individuals i α el vector de constants individuals. El model resultant és el següent:

$$D = \begin{pmatrix} \iota_T & \vec{0}_T & \cdots & \vec{0}_T \\ \vec{0}_T & \iota_T & \cdots & \vec{0}_T \\ \vdots & \vdots & \ddots & \vdots \\ \vec{0}_T & \vec{0}_T & \cdots & \iota_T \end{pmatrix}_{NT \times N}, \ \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} = \begin{pmatrix} \alpha + \mu_1 \\ \alpha + \mu_2 \\ \vdots \\ \alpha + \mu_N \end{pmatrix}_{N \times 1}$$

Si a aquesta especificació s'hi aplica l'estimador OLS, s'obté l'estimador LSDV. El problema de l'estimador LSDV és que, si N és elevat, hi haurà massa *dummies* individuals que s'han d'estimar i hi pot haver problemes en invertir grans matrius.

2) Within groups (WG): aquesta alternativa es basa a extreure la mitjana aritmètica de totes les variables. Per a la variable dependent:

$$\tilde{y}_{it} = y_{it} - \bar{y}_{i\cdot}, \qquad \text{on} \qquad \bar{y}_{i\cdot} = \frac{1}{T}\sum_{s=1}^{T} y_{is}$$

L'extracció de la mitjana de les variables elimina els efectes individuals, ja que:

$$\bar{\alpha}_i = \alpha_i \qquad \text{de manera que} \qquad \tilde{\alpha}_i = \alpha_i - \alpha_i = 0$$

Aquesta transformació elimina els efectes que no varien en el temps. En eliminar α_i, aquest ja no estarà correlacionat amb x_i, i això garanteix la condició d'ortogonalitat $E(x_{it}u_{it}) = 0$. Per a obtenir aquest estimador s'han d'aplicar algunes transformacions. Primer, creem una matriu de projecció de dimensió $T \times T$:

$$M_0 = I_T - \frac{1}{T}\iota\iota'$$

On ι és un vector d'uns de dimensió $T \times 1$. Aquesta matriu fa el següent:

$$M_0 Y_i = \begin{pmatrix} \tilde{y}_{i1} \\ \vdots \\ \tilde{y}_{iT} \end{pmatrix} = \begin{pmatrix} y_{i1} - y_{i\cdot} \\ \vdots \\ y_{iT} - y_{i\cdot} \end{pmatrix}_{T\times 1}$$

Generalitzem M_0 per aplicar-la a tot el model i extreure la mitjana de tots els individus:

$$M_d = I_{NT} - D(D'D)^{-1}D' = \begin{pmatrix} M_0 & 0_{T\times T} & \cdots & 0_{T\times T} \\ 0_{T\times T} & M_0 & \cdots & 0_{T\times T} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{T\times T} & 0_{T\times T} & \cdots & M_0 \end{pmatrix}_{NT\times NT}$$

Si multipliquem M_d per D, obtenim:

$$\begin{aligned} M_d D &= (I_{NT} - D(D'D)^{-1}D')D \\ &= I_{NT}D - D(D'D)^{-1}D'D \\ &= 0_{NT\times N}. \end{aligned}$$

Per acabar, premultipliquem l'especificació per M_d per eliminar efectes individuals:

$$M_d Y = M_d X\beta + M_d u$$

Apliquem OLS a aquest model transformat i obtenim l'estimador WG:

$$\hat{\beta}_{WG} = (X' M_d X)^{-1} X' M_d Y$$

Els efectes individuals estimats es poden recuperar aplicant la fórmula següent:

$$\hat{\alpha}_i = \hat{\alpha}_0 + \hat{\mu}_i = \bar{y}_{i\cdot} - \hat{\beta}'_{WG}\bar{x}_{i\cdot}$$

L'estimador d'efectes fixos posseeix les propietats següents:

1) L'estimador LSDV/WG és **consistent** tant per a $E(x_{it}\mu_i) = 0$ com per a $E(x_{it}\mu_i) \neq 0$, ja que la transformació elimina els efectes individuals.

2) Si es compleix que $E(x_{it}\mu_i) = 0$, l'estimador LSDV/WG serà **menys eficient** a mesura que $N \longrightarrow \infty$ amb T fixat. Això es deu al fet que estimar N constants implica una reducció de graus de llibertat.

3) Si es compleix que tots els regressors estan correlacionats amb μ_i, l'estimador LSDV/WG serà **eficient**.

4) *Limitació:* no es poden incloure variables que no variïn temporalment.

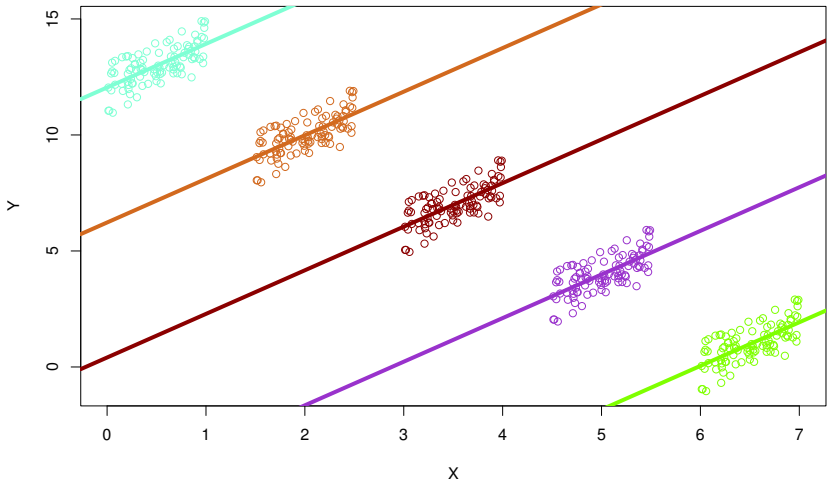
Reprenem l'exemple anterior de la relació entre renda i consum. Ara, assumim la hipòtesi d'efectes fixos:

$$C_{it} = \alpha + \mu_i + \beta R_{it} + u_{it}$$

En aquesta nova especificació, la presència de l'efecte individual μ_i pot mitigar l'absència de la variable patrimoni. Apliquem l'estimador WG/LSDV:

Variable	Coef.	Error est.	Estad. t	Valor-p
$\hat{\alpha}_1$	12,04	0,07	164,28	0,00
$\hat{\alpha}_2$	6,22	0,18	33,53	0,00
$\hat{\alpha}_3$	0,40	0,31	1,30	0,19
$\hat{\alpha}_4$	−5,41	0,44	−12,20	0,00
$\hat{\alpha}_5$	−11,23	0,57	−19,57	0,00
Renda ($\hat{\beta}$)	1,87	0,08	21,44	0,00
Estadístic F	459			0,00
R^2	0,48			

Veiem que, en aquest cas, tenim cinc rectes diferents, totes amb el mateix pendent però amb una constant que varia d'individu a individu.



3.2.3. Primeres diferències (FD)

En anglès es denomina *first differenced* OLS (FD). Aquest estimador es basa a calcular primeres diferències com a alternativa per a eliminar els efectes individuals. Si prenem el model inicial:

$$y_{it} = \alpha + \mu_i + x'_{it}\beta + u_{it}$$

Prenent primeres diferències s'eliminen α i μ_i, ja que aquests no varien amb el temps:

$$\Delta y_{it} = \Delta x'_{it}\beta + \Delta u_{it}$$

On $\Delta y_{it} = y_{it} - y_{i,t-1}$, per exemple. Si apliquem MCO a aquest model obtenim l'estimador $\hat{\beta}_{FD}$.

Les propietats d'aquest estimador són les següents:

1) L'estimador serà **consistent** si $E(\Delta x_{it}\Delta u_{it}) = 0$ (regressors exògens).

2) Si el terme u_{it} és persistent en el temps, Δu_{it} no tindrà correlació en sèrie i l'estimador FD serà **més eficient** que l'estimador WG/LSDV.

3) En canvi, si $\tilde{u}_{it}$ no mostra correlació en sèrie, FD serà **menys eficient** que WG/LSDV.

4) Com l'estimador WG/LSDV, no es poden incloure variables que no variïn temporalment.

3.2.4. Entre grups (BG)

En anglès, es denomina estimador *between groups* (BG). Aquest model inclou solament les mesures individuals de les variables:

$$\bar{y}_{i\cdot} = \bar{x}_i'\beta + \bar{\eta}_i + \bar{v}_{i\cdot}, \qquad i = 1, \ldots, N.$$

Aquest model solament inclou la dimensió individual, i no inclou informació temporal. Per això, no es pot estudiar la tendència dinàmica de les variables. Les seves principals propietats són que aquest estimador BG és **consistent** si $E(x_{it}\mu_i) = 0$. A més a més, és un estimador **no eficient**, ja que solament té $N - k$ graus de llibertat. Se sol usar com a pas previ abans d'efectuar altres estimacions per a comprovar la variabilitat de les observacions entre diferents grups.

3.2.5. Efectes aleatoris (RE/GLS)

En anglès es denomina *random effects generalized least squares* (RE/GLS). Aquest model és apropiat si s'extreuen N individus aleatòriament d'una població gran, i la mostra és representativa. Així, l'interès no és el conjunt de característiques de cada individu, sinó fer inferència sobre les característiques de la població.

L'especificació de partida és:

$$y_{it} = \alpha + x_{it}'\beta + u_{it}$$

$$u_{it} = \mu_i + \epsilon_i$$

L'efecte individual μ_i es pren com a aleatori (degut a l'atzar), i es modelitza com una variable (pertorbació) aleatòria. L'efecte μ_i és considerat una pertorbació aleatòria constant en el temps, homoscedàstica, no autocorrelacionada i amb mitjana zero:

$$\mu_i \sim iid(0, \sigma_\mu^2)$$

$$E(\mu_i\mu_j) = 0, \quad \forall i \neq j$$

$$E(\mu_i^2) = \sigma_\mu^2$$

$$E(\mu_i) = 0.$$

Similarment, obtenim el següent:

$$\epsilon_{it} \sim iid(0, \sigma_\epsilon^2)$$

$$E(\epsilon_{it}\epsilon_{js}) = 0, \quad \forall i \neq j,\ t \neq s$$

$$E(\epsilon_{it}^2) = \sigma_\epsilon^2.$$

$$E(\epsilon_i) = 0.$$

Per entendre la constucció d'aquest estimador, tornem enrere al model de regressió lineal:

$$y_i = x_i'\beta + u_i$$

Si s'analitza el terme de pertorbació u_i, s'assumeix que té mitjana condicional zero ($E(u_i|x_i) = 0$) i variància finita. La variància condicional del model serà $E(u_i^2|x_i) = \sigma_i^2$, i la matriu de variàncies i covariàncies de u és:

$$MVC(u) = E(uu') = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{pmatrix}_{N\times N}$$

Recordem que, perquè l'estimador MCO sigui eficient (mínima variància de l'estimació), aquesta matriu ha de ser **esfèrica**, és a dir:

1) Homoscedàstica: la variància de u no varia entre els elements de la mostra, de manera que $\sigma_i^2 = \sigma^2$ i els elements de la diagonal de $MVC(u)$ són idèntics.

2) No autocorrelacionada: si els elements fora de la diagonal no són nuls ($\sigma_{ij} \neq 0,\ \forall i \neq j$), el model de regressió està autocorrelacionat, i viceversa.

Si totes dues condicions es compleixen, la matriu $MVC(u)$ serà:

$$MVC(u) = E(uu') = \sigma^2 I_N$$

Essent I_N la matriu identitat de dimensió $N \times N$. En cas que $MVC(u)$ no sigui esfèrica, assumim que $MVC(u) = E(uu') = \Omega$. En aquest cas, MCO no incorpora l'estructura de l'error en l'estimació del model, i no és eficient. L'estimador eficient serà *mínims quadrats generalitzats* (GLS):

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

Com que Ω és desconegut, caldrà estimar-lo o imposar-li una estructura ($\hat{\Omega}$), i per després aplicar l'estimador per mínims quadrats generalitzats factibles (GLSF):

$$\hat{\beta}_{GLSF} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y$$

Tornant al model amb dades de panel, en el model d'efectes aleatoris la matriu $MVC(u)$ no és esfèrica.

$$E(u_{it}^2) = \sigma_\mu^2 + \sigma_\epsilon^2$$

$$E(u_{it}u_{is}) = \sigma_\mu^2, \quad \forall t \neq s.$$

Per a l'individu i, l'$MVC(u)$ serà:

$$\Omega_i = E(u_i u_i') = \sigma_\epsilon^2 I_T + \sigma_\mu^2 \iota\iota'$$

On ι és un vector d'uns $T \times 1$. La matriu Ω_i no serà esfèrica:

$$\Omega_i = \begin{pmatrix} \sigma_\mu^2 + \sigma_\epsilon^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\epsilon^2 & \cdots & \sigma_\mu^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\mu^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_\epsilon^2 \end{pmatrix}_{T\times T}$$

La matriu $MVC(u)$ per a tots els individus serà llavors:

$$\Omega = E(uu') = I_N \otimes \Omega_{T\times T},$$

Amb l'expressió analítica següent:

$$\Omega = \begin{pmatrix} \Omega_1 & 0_{T\times T} & \cdots & 0_{T\times T} \\ 0_{T\times T} & \Omega_2 & \cdots & 0_{T\times T} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{T\times T} & 0_{T\times T} & \cdots & \Omega_N \end{pmatrix}_{NT\times NT}.$$

Ω és una matriu diagonal en blocs que exhibeix correlació en sèrie en el temps.

Per a obtenir una estimació eficient del model, l'estructura del terme de pertorbació s'inclou en l'estimació mitjançant mínims quadrats generalitzats (GLS):

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.$$

Wansbeek i Kapteyn van idear un procediment per obtenir l'estimador, partint de la desviació de Ω^{-1} i $\Omega^{-1/2}$. Si es premultiplica el model:

$$Y^* = \Omega^{-1/2}Y$$

$$X^* = \Omega^{-1/2}X$$

$$\epsilon^* = \Omega^{-1/2}\epsilon,$$

S'obté l'estimador RE/GLS aplicant mínims quadrats ordinaris al model:

$$\begin{aligned}\hat{\beta}_{GLS} &= (X^{*'}X^*)^{-1}X^{*'}Y^* \\ &= (X'\Omega^{-1/2}\Omega^{-1/2}X)^{-1}X'\Omega^{-1/2}\Omega^{-1/2}Y \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.\end{aligned}$$

Aquest prodeciment es denomina *θ-diferenciació*, i per a aplicar-lo s'ha de definir la següent matriu de variàncies i covariàncies (MVC) individual:

$$\Omega_i^{-1/2} = I_T - \frac{\theta}{T}\iota\iota', \qquad \theta = 1 - \frac{\sigma_\epsilon}{\sqrt{T\sigma_\mu^2 + \sigma_\epsilon^2}},$$

On ι és un vector d'uns $T \times 1$.

La MVC global es deriva llavors de la manera següent:

$$\Omega^{-1/2} = I_N \otimes \Omega_i^{-1/2}.$$

Vegem com la transformació afecta les variables:

$$Y_i^* = \Omega_i^{-1/2}Y_i = \begin{pmatrix} y_{i1} - \theta\bar{y}_{i\cdot} \\ \vdots \\ y_{iT} - \theta\bar{y}_{i\cdot} \end{pmatrix}_{T\times 1}.$$

Depenent dels valors de les variàncies σ_μ^2 i σ_ϵ^2, el valor θ variarà, cosa que té implicacions respecte a l'estimació. No obstant això, els valors d'aquestes variàncies, i per tant la matriu Ω, són desconegudes. Per això, s'ha d'estimar $\hat{\sigma}_\mu^2$ i $\hat{\sigma}_\epsilon^2$, i per a fer-ho hi ha diversos procediments. En aquest cas, l'estimador es denomina mínims quadrats generalitzats factibles (FGLS).

Les propietats de l'estimador d'efectes aleatoris són les següents:

1) Requereix que les variables explicatives no estiguin correlacionades amb els efectes individuals: $E(x_{it}\mu_i) = 0$. En aquest cas, l'estimador RE/GLS serà **consistent** i **eficient**.

2) Si $E(x_{it}\mu_i) \neq 0$, RE/GLS és **inconsistent** a mesura que $N \longrightarrow \infty$ amb T fixat.

3) Si no hi ha efectes individuals, de manera que $\sigma_\mu^2 = 0$, llavors $\theta = 0$, i els estimadors RE/GLS i OLS coincidiran i tots dos seran **eficients**. Això succeeix perquè aplicar GLS a un model amb MVC esfèrica equival a aplicar l'estimador de mínims quadrats OLS.

4) A mesura que $T \longrightarrow \infty$, $\theta \longrightarrow 1$ i RE/GLS tendirà a coincidir amb l'estimador WG.

5) A diferència de l'estimador d'efectes fixos, aquest estimador permet incloure regressors que no variïn temporalment.

3.2.6. Coeficients variables

En anglès es denomina *variable coefficients model*. Aquest model relaxa la hipòtesi que els paràmetres del model són comuns ($\beta_{ij} = \beta$). D'aquesta manera, el model de coeficients variables aporta més flexibilitat i permet estimar un coeficient per a cada individu ($\hat{\beta}_i$) i/o per a cada període ($\hat{\beta}_t$), a més de la constant ($\alpha_i,\ \alpha_t$).

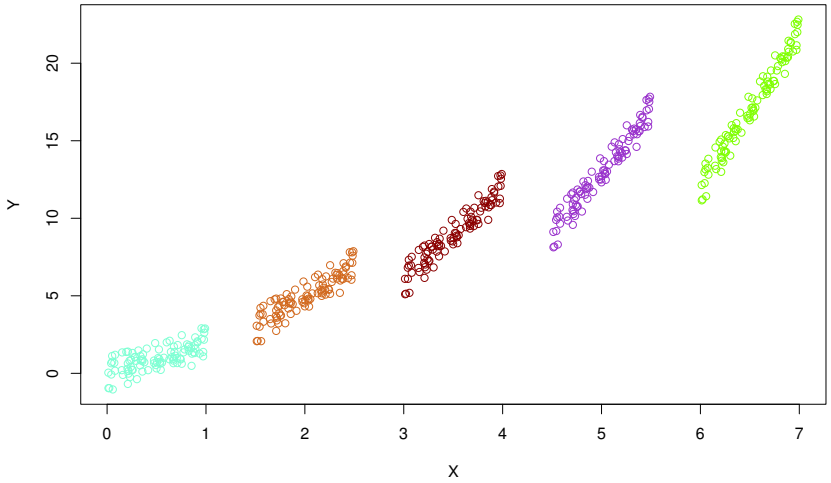
El model de coeficients variables pot ser estimat amb efectes fixos o efectes aleatoris:

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}.$$

Com a exemple, reprenem el model de renda i consum analitzat anteriorment:

$$C_{it} = \alpha + \beta R_{it} + u_{it}$$

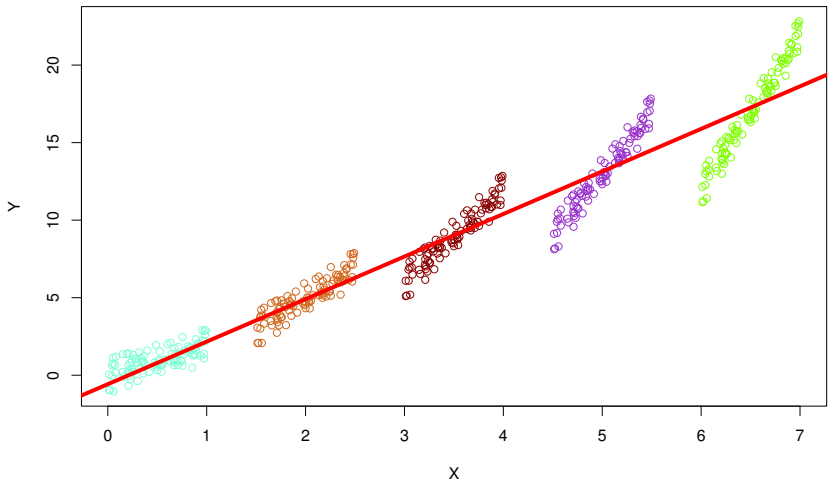
Suposem que ens trobem davant el següent diagrama de dispersió de les dues variables:



Si estimem el model amb *pooled OLS*, obtenim el resultat següent:

Variable	Coef.	Error est.	Estad. t	Valor-p
Constante $\hat{\alpha}$	$-0,57$	$0,11$	$-4,87$	$0,00$
Renda $\hat{\beta}$	2.74	$0,03$	$94,97$	$0,00$
Estadístic F	9019			$0,00$
R^2	$0,94$			

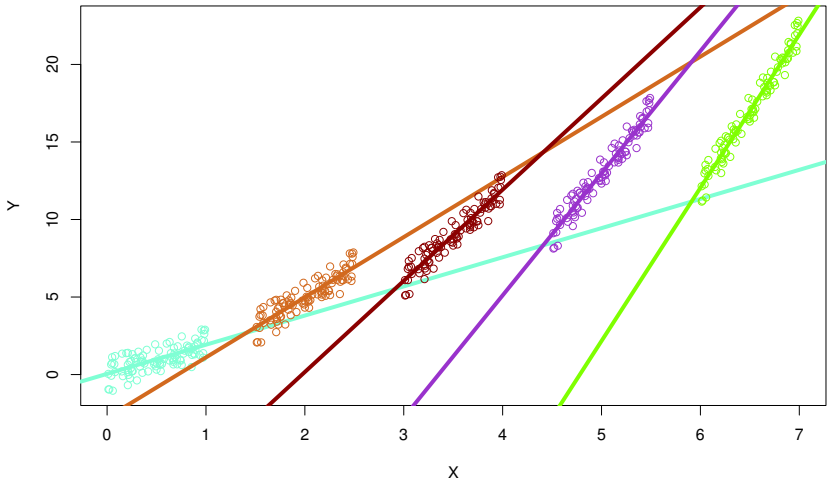
Es pot comprovar que l'ajust del model és bastant bo, i els coeficients estimats són significatius. Vegem ara l'ajust de la recta estimada sobre les observacions.



Es pot comprovar que la recta estimada no s'ajusta als diferents individus de la mostra. Si estiméssim un model de coeficients variables amb efectes fixos, obtindríem el resultat següent:

i	α_i	β_i
1	0,04	1,87
2	−2,77	3,87
3	−11,59	5,87
4	−26,41	7,87
5	−47,23	9,87
R^2	0,99	

Comprovem gràficament que l'ajust de l'estimació a les observacions ha millorat amb aquesta nova estimació.



3.2.7. Mètode generalitzat dels moments (GMM)

En anglès es denomina *generalized method of moments (GMM)*. En dades de panel, és molt freqüent l'anàlisi d'equacions dinàmiques del tipus:

$$y_{it} = \rho y_{i,t-1} + \beta x_{it} + \mu_i + u_{it}$$

Els models que incorporen un retard de la variable explicada com a regressor pateixen d'endogenitat. Això és, es trenca el supòsit d'exogenitat dels regressors, i a més passa que $E(y_{i,t-1}u_{it} \neq 0)$. Això succeeix també quan es prenen les primeres diferències per a eliminar els efectes individuals.

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \beta \Delta x_{it} + \Delta u_{it}$$

La solució és utilitzar variables instrumentals i l'estimador GMM per a eliminar el biaix causat per l'endogenitat.

3.3. Inferència

A l'hora d'especificar i estimar un model economètric de dades de panel, ens trobem davant moltes possibilitats. Per a encertar en la tria de l'especificació i del mètode d'estimació hi ha diferents contrastos, que ens poden ajudar a fer una elecció basada en criteris estadístics. En aquesta secció s'introdueixen solament els tres contrastos més coneguts i usats.

1) Contrast d'efectes fixos: aquest contrast es basa en la hipòtesi que els termes constants són tots iguals. Si aquesta hipòtesi és certa, no existirien els efectes individuals. La hipòtesi del contrast és la següent:

$$H_0 : \alpha_i = \alpha, \quad \forall i = 1, \ldots, N$$

$$H_1 : \alpha_i \neq \alpha, \quad \forall i = 1, \ldots, N$$

L'estadístic F del contrast és:

$$F = \frac{(SCR_{OLS} - SCR_{WG})/(N-1)}{SCR_{WG}/(NT-N-K)} \sim F_{(N-1),(NT-N-K)}$$

On SCR és la suma dels quadrats dels residus $\sum \hat{u}_i^2$. Si l'ajust del model WG millora OLS perquè hi ha efectes individuals, la diferència $SCR_{OLS} - SCR_{WG}$ serà més gran i F caurà en la regió crítica (rebuig de H_0).

2) Contrast d'efectes aleatoris: es tracta d'un contrast de multiplicadors de Lagrange, i es basa en els residus del model OLS. La hipòtesi del contrast és la següent:

$$H_0 : \sigma_\mu^2 = 0 \implies corr(u_{it}u_{is}) = 0$$

$$H_1 : \sigma_\mu^2 \neq 0 \implies corr(u_{it}u_{is}) \neq 0$$

L'estadístic és el següent:

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} (\sum_{t=1}^{T} \hat{u}_{it})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^2} - 1 \right]^2 \sim \chi_1^2$$

Si hi ha variació intragrups (*within*), és a dir si $\sigma_\mu^2 > 0$, l'estadístic LM pren un valor alt i cau en la zona de rebuig de H_0. En aquest cas, hi ha efectes aleatoris.

3) Test de Hausman: aquest test s'usa per a decidir entre els estimadors d'efectes fixos i aleatoris (WG/LSDV vs. RE/GLS). S'ha de recordar el següent:

$E(x_{it}\mu_i) = 0$: WG i RE són consistents, i RE és l'estimador eficient (mínima variància).

$E(x_{it}\mu_i) \neq 0$: RE és inconsistent, i WG és consistent.

El test de Hausman es basa en el fet que si $E(x_{it}\mu_i) = 0$, WG i RE haurien de ser similars. Per això, el test determinarà si hi ha o no autocorrelació. La hipòtesi del contrast és la següent:

$$H_0 : E(x_{it}\mu_i) = 0$$

$$H_1 : E(x_{it}\mu_i) \neq 0$$

L'estadístic del contrast adquireix la forma següent:

$$\hat{q} = \hat{\beta}_{WG} - \hat{\beta}_{RE}$$

$$H = \hat{q}'[avar(\hat{q})]^{-1}\hat{q} \sim \chi^2_k$$

On $avar(\hat{q})$ és la variància asimptòtica de $\hat{q}$. La decisió final és la següent:

No-rebuig de H_0: l'estimador d'efectes aleatoris és preferit, ja que $E(x_{it}\mu_i) = 0$.

Rebuig de H_0: l'estimador d'efectes fixos (WG) és preferit, ja que $E(x_{it}\mu_i) \neq 0$.

3.4. Aplicació pràctica amb R

En aquesta secció revisem l'anàlisi economètrica de dades de panel amb R. Per il·lustrar-la, considerem un estudi de la productivitat de les manufactures espanyoles per a diferents sectors i anys. En aquest sentit, es consideren les variables següents:

Y : valor afegit brut.

L : quantitat de factor treball.

K : estoc de capital.

Aquestes variables estan disponibles per a $i = 1, \ldots, N$ sectors i per a $j = 1, \ldots, T$ anys. Com que $N = 11$ sectors i $T = 17$ anys (de 1980 a 1996), la dimensió total del panel és de $NT = 187$ observacions.

El punt de partida és una especificació Cobb-Douglas de la forma següent:

$$Y = AL^{\beta_L}K^{\beta_K}$$

Per a estimar empíricament aquest model, s'opta per una transformació logarítmica:

$$logY_{it} = \alpha + \beta_L logL_{it} + \beta_K logK_{it} + u_{it}$$

Començarem especificant el directori de treball (on són les dades) mitjançant l'ordre `setwd()` i carregant les biblioteques bàsiques:

És fonamental que les biblioteques s'han d'instal·lar abans de poder carregar-les. Recordem que la instrucció per a carregar-les és `install.packages()`.

```
> library(plm)
> library(car)
> library(gplots)
```

La biblioteca`plm` es denomina *Linear Models for Panel Data*, i és la biblioteca de referència per a l'estimació d'aquest tipus de models. La biblioteca `car` es denomina *Companion to Applied Regression*, i la biblioteca `gplots` és *Various R programming tools for plotting data*. Aquestes dues últimes les carreguem per fer alguns gràfics descriptius previs multidimensionals, això és, amb dades amb dues dimensions (individus i temps). Seguidament, llegim el document amb les dades per obtenir la base de dades:

```
> datos<-read.delim2("datos.txt",header=TRUE,dec=",")
```

Per a una primera aproximació a les dades, farem el següent:

```
> summary(datos)

      year          sector          Y
 Min.   :1980   Min.   : 1   Min.   : 213814
 1st Qu.:1984   1st Qu.: 3   1st Qu.: 449129
 Median :1988   Median : 6   Median : 642399
 Mean   :1988   Mean   : 6   Mean   : 705955
 3rd Qu.:1992   3rd Qu.: 9   3rd Qu.: 838315
 Max.   :1996   Max.   :11   Max.   :1986698
       L                K
 Min.   : 80574   Min.   : 1492221
 1st Qu.:111936   1st Qu.: 3141607
 Median :142229   Median : 4256334
 Mean   :177993   Mean   : 5502817
 3rd Qu.:230264   3rd Qu.: 7285621
 Max.   :384167   Max.   :14903525
```

La funció `head()` ens ensenya la capçalera de les dades:

```
> head(datos)

  year sector      Y      L        K
1 1980      1 718333 205073 13635632
2 1981      1 687133 193424 13583555
3 1982      1 597825 173109 13290383
4 1983      1 591872 158456 13031579
5 1984      1 567590 152955 12821882
6 1985      1 545167 142229 12908390
```

L'objecte creat `datos` és del tipus `"data.frame"`, és a dir, una base de dades. Per efectuar l'anàlisi de dades de panel, hem de crear un nou objecte del tipus `"pdata.frame"`, que incorpora informació sobre les dues dimensions que conformen el panel, és a dir, la individual i la temporal:

```
> datos.pd<-pdata.frame(datos,index=c("sector","year"))
```

Visualitzem ara la capçalera de la nova base de dades.

```
> head(datos.pd)

       year sector      Y      L        K
1-1980 1980      1 718333 205073 13635632
1-1981 1981      1 687133 193424 13583555
1-1982 1982      1 597825 173109 13290383
1-1983 1983      1 591872 158456 13031579
1-1984 1984      1 567590 152955 12821882
1-1985 1985      1 545167 142229 12908390
```

A continuació, introduïm les variables incloses en la base de dades en l'espai de treball, per poder operar-hi directament.

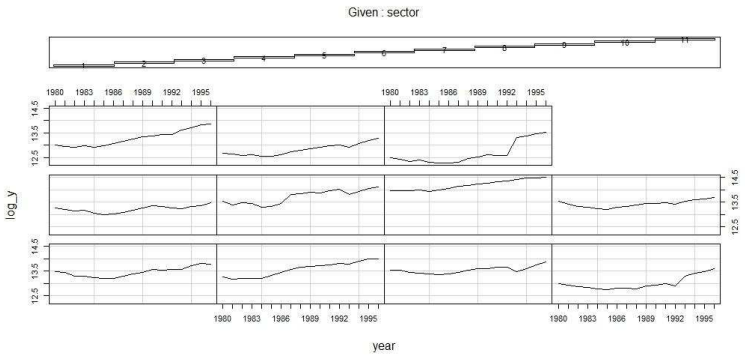
```
> attach(datos.pd)
```

Per estimar el model especificat, abans de res crearem les tres variables en logaritmes:

```
> log_y<-log(Y)
> log_l<-log(L)
> log_k<-log(K)
```

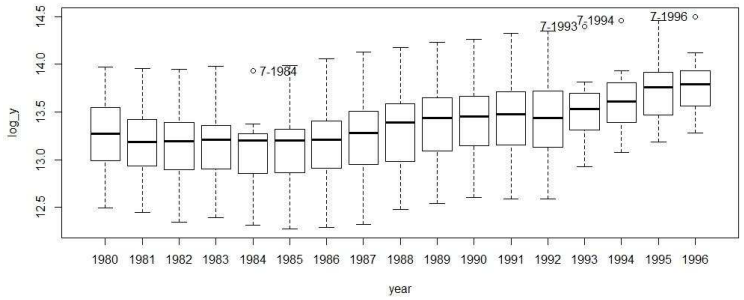
Una funció que ens ajuda a entendre l'estructura bidimensional de les dades és `coplot()`, ja que és una eina gràfica molt intuïtiva. Analitzem l'estructura de la variable renda:

```
> coplot(log_y year|sector,type="l",data=datos.pd)
```



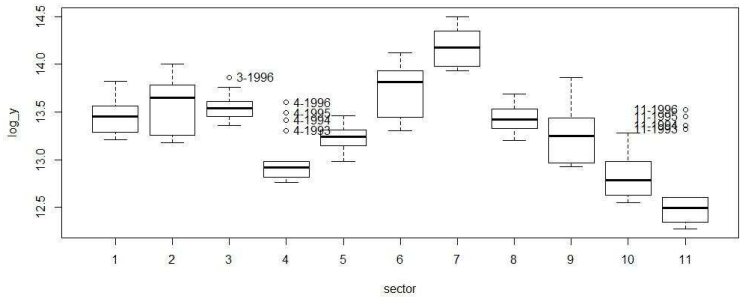
Com veiem, l'evolució temporal de $log(Y)$ és positiva per a tots els sectors, però amb efectes específics en cada sector. Una altra opció interessant seria un gràfic en què es pugui veure, per a cada any, la distribució de $log(Y)$ per als diferents sectors. Ho farem amb la instrucció següent:

```
> scatterplot(log_y year|sector,data=datos.pd)
```



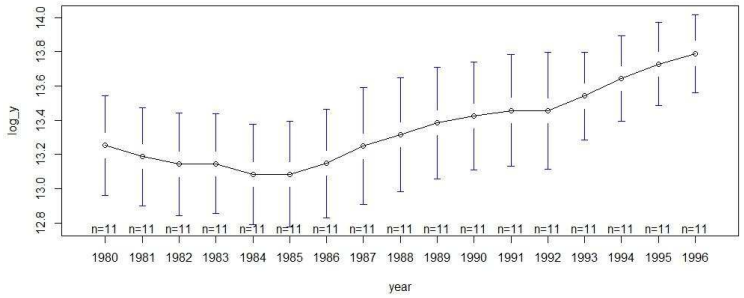
El mateix gràfic però aplicat a diferents sectors permet veure com la mitjana temporal de $log(Y)$ és diferent per a cada sector, de manera que, intuïtivament, és concebible que hi pugui haver efectes individuals.

```
> scatterplot(log_y sector|year,data=datos.pd)
```

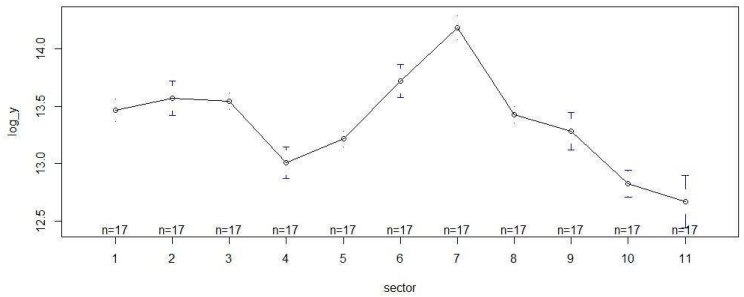


Una aproximació alternativa és usar la funció `plotmeans()`, que produeix resultats similars a la funció anterior. Si apliquem aquesta funció a totes dues dimensions (sector i any), comprovem que els efectes sectorials són significatius.

```
> plotmeans(log_y year)
```




```
> plotmeans(log_y sector)
```



A l'hora d'especificar i estimar el model de regressió, el paquet `plm` permet treballar amb tres efectes: l'individual (`effect=ïndividual"`), el temporal (`effect="time"`) i el conjunt de tots dos efectes simultàniament (`effect="twoways"`). Si no l'especifiquem, R pren per defecte efectes individuals. A partir d'ara, per simplicitat, elabora-

rem l'exemple assumint solament un efecte individual (sectorial), encara que l'anàlisi és fàcilment generalitzable. Per això, l'especificació que triem és la següent:

$$logY_{it} = \alpha + \beta_L logL_{it} + \beta_K logK_{it} + \mu_i + u_{it}$$

El primer pas és introduir l'especificació del model, és a dir, la fórmula de la regressió que s'estimarà. Ho farem creant un objecte del tipus `"formula"`.

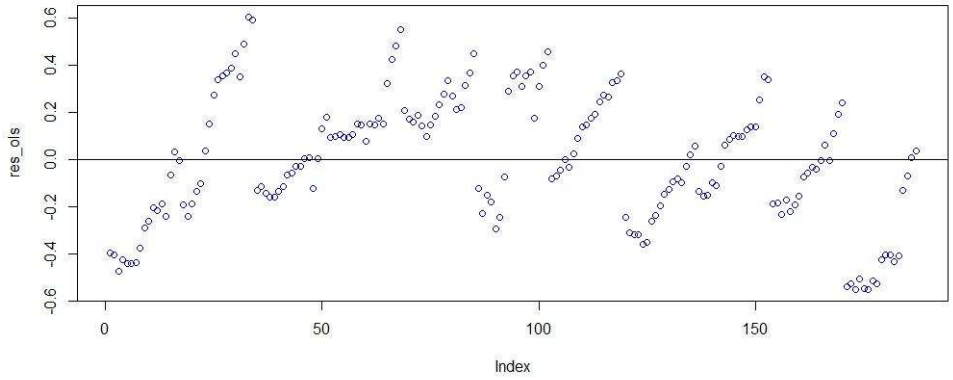
```
> eq1<-log_y~log_l+log_k
```

El primer pas serà efectuar una regressió *pooled OLS*, és a dir, mínims quadrats ordinaris ignorant els efectes individuals. El resultat és el següent:

```
> m_ols<-plm(eq1,data=datos.pd,model="pooling")
> summary(m_ols)

Oneway (individual) effect Pooling Model

Call:
plm(formula = eq1, data = datos.pd, model = "pooling")

Balanced Panel: n=11, T=17, N=187

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.55400 -0.18300 -0.00176  0.18100  0.60600

Coefficients :
            Estimate Std. Error t-value Pr(>t)
(Intercept) 0.871465   0.618935  1.4080   0.1608
log_l       0.578346   0.051686 11.1897   <2e-16 ***
log_k       0.361428   0.038264  9.4456   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    42.488
Residual Sum of Squares: 13.223
R-Squared      :  0.68878
      Adj. R-Squared :  0.67773
F-statistic: 203.611 on 2 and 184 DF, p-value: < 2.22e-16
```

Veiem que l'ajust és raonablement bo, i els coeficients estimats són estadísticament significatius. Una bona eina d'anàlisi del model és una primera inspecció visual dels residus de la regressió.

```
> res_ols<-m_ols[["residuals"]]
> plot(res_ols,col="blue")
> abline(h=0)
```



L'estructura del residu, en el gràfic, revela que el comportament no sembla aleatori, ja que els efectes individuals no s'han tingut en compte i aquests es manifiesten en el residu de la regressió.

El pas següent és estimar el model mitjançant l'estimador d'efectes fixos (*within groups*).

```
> m_wg<-plm(eq1,data=datos.pd,model="within",effect="individual
    ")
```

Abans d'analitzar els resultats, ens hem de preguntar: els efectes fixos són significatius? És a dir, hi ha una constant comuna per a tots els individus, o és diferent? Aquí, dos contrastos són d'utilitat. El primer és un contrast del tipus F, que avalua la hipòtesi nul·la que $\alpha_i = \alpha$. A R:

```
> pFtest(m_wg,m_ols)

        F test for individual effects

data:  eq1
F = 46.1892, df1 = 10, df2 = 174, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Segons el contrast F, la constant és diferent per a cada individu. Una alternativa és fer un test de multiplicadors de Lagrange, que avalua la mateixa hipòtesi nul·la, que com veiem a continuació dóna el mateix resultat.

```
> plmtest(m_ols,effect="individual")

        Lagrange Multiplier Test - (Honda)

data:  eq1
normal = 21.0093, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Arribats a aquest punt, vegem el resultat de l'estimació amb efectes fixos.

```
> summary(m_wg)

Oneway (individual) effect Within Model

Call:
plm(formula = eq1, data = datos.pd, effect = "individual",
    model = "within")

Balanced Panel: n=11, T=17, N=187

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.49900 -0.06520  0.00527  0.06550  0.37400

Coefficients :
      Estimate Std. Error t-value  Pr(>t)
log_l 0.757130   0.078966  9.5881 < 2.2e-16 ***
log_k 0.999751   0.074056 13.4999 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    11.582
Residual Sum of Squares: 3.6183
R-Squared      :  0.68758
      Adj. R-Squared :  0.63978
F-statistic: 191.475 on 2 and 174 DF, p-value: < 2.22e-16
```

També tenim l'opció d'extreure les constants individuals.

```
> summary(fixef(m_wg))

   Estimate Std. Error t-value  Pr(>t)
1  -12.0523     1.3056 -9.2313 < 2.2e-16 ***
2  -11.0916     1.2613 -8.7937 < 2.2e-16 ***
3  -11.2379     1.2672 -8.8682 < 2.2e-16 ***
4  -10.2121     1.1875 -8.5998 < 2.2e-16 ***
5  -10.4387     1.2092 -8.6325 < 2.2e-16 ***
6  -11.1690     1.2722 -8.7792 < 2.2e-16 ***
7  -11.4950     1.3128 -8.7559 < 2.2e-16 ***
8  -11.2453     1.2622 -8.9093 < 2.2e-16 ***
9  -10.9611     1.2394 -8.8437 < 2.2e-16 ***
10 -10.7709     1.2063 -8.9288 < 2.2e-16 ***
11 -11.0980     1.2150 -9.1341 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Una manera alternativa d'eliminar els efectes fixos consisteix a aplicar l'estimador en primeres diferències.

```
> m_fd<-plm(eq1,data=datos.pd,model="fd")
```

Quin estimador és més efectiu per a eliminar els efectes individuals, *wg* o *fd*? La resposta està en si el terme de pertorbació u_{it} mostra correlació en sèrie. Per a trobar una resposta a aquesta pregunta hi ha un test que es basa a estimar el model següent:

$$\hat{u}_{it} = \rho \hat{u}_{i,t-1} + e_{it}$$

Si $\hat{\rho} \rightarrow 0$, llavors l'estimador més adequat serà efectes fixos (*wg*), mentre que si $\hat{\rho} \rightarrow 1$, llavors tindrà sentit aplicar primeres diferències per a així eliminar (a més dels efectes fixos) la correlació de la pertorbació ρ. El test que s'ha d'aplicar és el *test de primeres diferències de Wooldridge*, que en una de les seves formes contrasta (sota H_0) si $Cor(u_{it}u_{i,t-1}) = 0$.

```
> pwfdtest(eq1,data=datos.pd,h0="fe")

        Wooldridge's first-difference test for serial
        correlation in panels

data:  plm.model
chisq = 57.9424, p-value = 2.699e-14
alternative hypothesis: serial correlation in original errors
```

Aquest resultat sembla indicar que $Cor(u_{it}u_{i,t-1}) \neq 0$. Això és una indicació per a usar el model *fd*. En el model de primeres diferències el terme de pertorbació passaria a ser

$$e_{it} = u_{it} - u_{i,t-1}.$$

Un segon contrast confirmarà si $Cor(e_{it}e_{i,t-1}) = 0$, cosa que donaria validesa a l'opció del model *fd*:

```
> pwfdtest(eq1,data=datos.pd,h0="fd")

        Wooldridge's first-difference test for serial
        correlation in panels

data:  plm.model
chisq = 0.0612, p-value = 0.8046
alternative hypothesis: serial correlation in differenced
    errors
```

Efectivament, l'estimador preferit per a eliminar (*wipe out*) els efectes individuals és *fd*.

L'estimador entre grups *between groups* s'aplica a les mesures aritmètiques de les variables, amb la qual cosa és un estimador que no aprofita l'estructura de dues dimensions. Si fem una regressió prenent les mitjanes dels sectors, obtenim:

```
> m_bg_i<-plm(eq1,data=datos.pd,model="between",effect="
    individual")
> summary(m_bg_i)

Oneway (individual) effect Between Model

Call:
plm(formula = eq1, data = datos.pd, effect = "individual",
    model = "between")

Balanced Panel: n=11, T=17, N=187

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-0.4080 -0.1320  0.0612  0.1650  0.2220
```

```
Coefficients :
            Estimate Std. Error t-value Pr(>t)
(Intercept)  1.70898    2.31037  0.7397  0.48062
log_l        0.56305    0.19817  2.8413  0.02177 *
log_k        0.31881    0.14357  2.2206  0.05713 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1.818
Residual Sum of Squares: 0.43502
R-Squared      :  0.76072
      Adj. R-Squared :  0.55325
F-statistic: 12.7169 on 2 and 8 DF, p-value: 0.0032781
```

En canvi, prenent les mitjanes dels anys (és a dir, agregant sectors):

```
> m_bg_t<-plm(eq1,data=datos.pd,model="between",effect="time")
> summary(m_bg_t)

Oneway (time) effect Between Model

Call:
plm(formula = eq1, data = datos.pd, effect = "time", model = "
    between")

Balanced Panel: n=11, T=17, N=187

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.10100 -0.04540 -0.00802  0.05150  0.10200

Coefficients :
             Estimate Std. Error t-value  Pr(>t)
(Intercept) -14.37941    2.56102 -5.6147 6.378e-05 ***
log_l         0.39246    0.21274  1.8448   0.08633 .
log_k         1.50029    0.13899 10.7942 3.596e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    0.80441
Residual Sum of Squares: 0.06308
R-Squared      :  0.92158
      Adj. R-Squared :  0.75895
F-statistic: 82.2647 on 2 and 14 DF, p-value: 1.8236e-08
```

Com veiem, l'estimació és sensiblement diferent, cosa que ens obliga a ser molt rigorosos en triar el model que s'ha d'estimar i en interpretar-ne els resultats.

L'estimador que ens queda és el d'efectes aleatoris (*random effects*), que serà consistent sempre que els efectes individuals (que són part de la pertorbació segons el model *random effects*) no estiguin correlacionats amb les variables explicatives. Primer estimem el model i a continuació efectuem el test de Hausman, que contrasta la hipòtesi nul·la que $E(x_{ij}\mu_i) = 0$.

```
> m_re<-plm(eq1,data=datos.pd,model="random")
> phtest(m_wg,m_re)

        Hausman Test

data:  eq1
chisq = 49.5461, df = 2, p-value = 1.743e-11
alternative hypothesis: one model is inconsistent
```

El test de Hausman sembla indicar el rebuig de H_0, amb la qual cosa el model *random effects* no és consistent i l'estimador *within effects* és consistent.

Un últim model estàtic és el de coeficients variables, que assumeix que $\beta_i \neq \beta$. El contrast efectuat amb la funció `pooltest()` ens indicarà si realment hi ha uns paràmetres específics per a cada sector o si els paràmetres són comuns.

```
> m_vc<-pvcm(eq1,data=datos.pd,model="within",effect="
    individual")
> pooltest(m_ols,m_vc)

        F statistic

data:  eq1
F = 20.3495, df1 = 30, df2 = 154, p-value < 2.2e-16
alternative hypothesis: unstability
```

El resultat del test sembla indicar l'existència de paràmetres individuals. Aplicant la funció `summary()` a l'estimació n'obtenim un resum.

```
> summary(m_vc)

Oneway (individual) effect No-pooling model

Call:
pvcm(formula = eq1, data = datos.pd, effect = "individual",
    model = "within")
```

```
Balanced Panel: n=11, T=17, N=187

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.46650 -0.04995 -0.01030  0.00000  0.05380  0.38560

Coefficients:
  (Intercept)          log_l             log_k
 Min.   :-38.606   Min.   :-1.7687   Min.   :0.6171
 1st Qu.:-11.329   1st Qu.: 0.1800   1st Qu.:0.8511
 Median : -9.663   Median : 0.5613   Median :1.0105
 Mean   : -9.841   Mean   : 0.2757   Mean   :1.2908
 3rd Qu.: -5.536   3rd Qu.: 0.6973   3rd Qu.:1.4670
 Max.   : 25.708   Max.   : 1.0220   Max.   :2.7434

Total Sum of Squares: 1206.2
Residual Sum of Squares: 2.6637
Multiple R-Squared: 0.99779
```

Per obtenir una estimació de tots els coeficients individuals, farem:

```
> summary(m_vc)[["coefficients"]]


   (Intercept)      log_l     log_k
1   -38.605905  0.5810760 2.7434385
2    -5.465967 -0.1025732 1.2826383
3    -9.663154  0.6943287 0.9480257
4   -11.426971  1.0219711 0.8702287
5    -5.605258  0.7002193 0.7183270
6    25.708400 -1.7687099 0.6170522
7    -4.221188 -0.6289064 1.6513321
8   -29.509106  0.5311384 2.3891937
9    -7.662915  0.4625070 1.0105400
10  -10.569548  0.5612735 1.1359098
11  -11.231134  0.9799564 0.8320530
```

Interpretant aquests resultats es podria inferir que la contribució dels factors K i L sobre el valor afegit Y varia segons el sector.

Bibliografia

Artís Ortuño, M.; del Barrio Castro, T.; Clar López, M.; Guillén Estany, M.; Suriñach Caralt, J. (2011). *Econometría*. Barcelona. Material didàctic UOC.

Liviano Solís, D.; Pujol Jover, M. (2013). *Matemáticas y Estadística con R*. Barcelona. Material didàctic UOC.